# 24th Annual National Conference on Managing Environmental Quality Systems

**8:30 – 12:00 TUESDAY, APRIL 12$^{TH}$ - A.M. Stockholder Meetings**

**12:00 – 4:30 TUESDAY, APRIL 12$^{TH}$**
**Opening Plenary** (Salons A-H)

- Opening Address
  - Reggie Cheatham, Director, OEI Quality Staff, EPA
  - Linda Travers, Principal Deputy Assistant Administrator, OEI, EPA
- Invited Speakers
  - Tom Huetteman, Deputy Assistant Regional Administrator, EPA Region 9
  - John Robertus, Executive Officer of San Diego Regional Water Quality Control Board, Region 9
- Keynote Address
  - Thomas Redman, President, Navesink Consulting Group
- Panel Sessions
- **Value of the Data Quality Act—Perspectives from OMB, Industry, and EPA (VDQA)**
  - Nancy Beck, OMB
  - Jamie Conrad, American Chemistry Council
  - Reggie Cheatham, Director, OEI Quality Staff, EPA
- **Wadeable Streams: Assessing the Quality of the Nation's Streams (WS)**
  - Margo Hunt, Panel Moderator
  - Mike Shapiro, Deputy Assistant Administrator, Office of Water
  - Steve Paulsen, Research Biologist, ORD

**8:30 – 10:00 WEDNESDAY, APRIL 13$^{TH}$**
**Environmental Measures (EM)** (Salons A-C) *Chair: L. Bradley, EPA*

- Data Error Reduction by Automation throughout the Data Workflow Process (A. Gray, EarthSoft, Inc.)
- Analytical Approaches to Meeting New Notification Levels for Organic Contaminants in Calif. (D.Wijekoon, Calif. DHS)
- Streamlining Data Management and Communications for the Former Walker AFB Project (R. Amano, Lab Data Consultants, Inc.)

**Quality System Implementation in the Great Lakes Program (QSI-GLP)** (Salon D) *Chair: M. Cusanelli, EPA*

- GLNPO's Quality System Implementation for the New "Great Lakes Legacy Act for Sediment Remediation"(L. Blume, EPA)
- Black Lagoon Quality Plan Approval by GLNPO, MDEQ, ERRS, and USACE (J. Doan, Environmental Quality Management, Inc.)
- Remediation of the Black Lagoon Trenton Channel . . . Postdredging Sampling & Residuals Analysis (J. Schofield, CSC)

**Quality Systems Models (QSM)** (Salons F-H) *Chair: G. Johnson, EPA*

- Improving E4 Quality System Effectiveness by Using ISO 9001: 2000 Process Controls (C. Hedin, Shaw Environmental)

**Applications of Novel Techniques to Environmental Problems (ANTEP)** (Salon E) *Chair: B. Nussbaum, EPA*

- On Some Applications of Ranked Set Sampling (B. Sinha, University of Maryland)
- Combining Data from Many Sources to Establish Chromium Emission Standards (N. Neerchal, University of Maryland)
- Estimating Error Rates in EPA Databases for Auditing Purposes (H. Lacayo, Jr., EPA)
- Spatial Population Partitioning Using Voronoi Diagrams For Environmental Data Analysis (A. Singh, UNLV)

**Ambient Air Session I (Sierra 5&6)** *Chair: M.Papp, EPA*

- Changes and Improvements in the Ambient Air Quality Monitoring Program Quality System (M. Papp, EPA)
- Guidance for a New Era of Ambient Air Monitoring (A. Kelley, Hamilton County DES)
- Environmental Monitoring QA in Indian Country (M. Ronca-Battista, Northern Arizona University)
- Scalable QAPP IT Solution for Air Monitoring Programs (C. Drouin, Lake Environmental Software)

**10:30 – 12:00 WEDNESDAY, APRIL 13TH**

**Environmental Laboratory Quality Systems (ELQS)** (Salons A-C) *Chair: L. Bradley, EPA*

- A Harmonized National Accreditation Standard: The Next Step for INELA Field Activities (D. Thomas, Professional Service Industries, Inc.)
- Development of a Comprehensive Quality Standard for Environmental Laboratory Accreditation (J. Parr, INELA)
- Advanced Tracking of Laboratory PT Performance and Certification Status with Integrated Electronic NELAC-Style Auditing Software (T. Fitzpatrick, Lab Data Consultants, Inc.)

**Performance Metrics (PM)** (Salon D) *Chair: L. Doucet, EPA*

- Formulating Quality Management Metrics for a State Program in an Environmental Performance Partnership Agreement (P. Mundy, EPA)
- How Good Is "How Good Is?" (Measuring QA) (M. Kantz, EPA)
- Performance-Based Management (J. Santillan, US Air Force)

**Quality Assurance Plan Guidance Initiatives (QAPGI)** (Salons F-H) *Chair: A. Batterman, EPA*

- A CD-ROM Based QAPP Preparation Tool for Tribes (D. Taylor, EPA)
- Military Munitions Response Program Quality Plans (J. Sikes, U.S. Army)

**Ask a Statistician: Panel Discussion** (Salon E) *Moderator: B. Nussbaum, EPA* Panelists:

- Mike Flynn, Director, Office of Information Analysis and Access, OEI, EPA
- Reggie Cheatham, Director, Quality Staff, OEI, EPA
- Tom Curran, Chief Information Officer, OAQPS, EPA
- Diane Harris, Quality Office, Region 7, EPA
- Bill Hunt, Visiting Senior Scientist, North Carolina State University (NCSU)
- Rick Linthurst, OIG, EPA

**Ambient Air Session II** (Sierra 5&6) *Chair: M. Papp, EPA*

- National Air Toxics QA System and Results of the QA Assessment (D. Mikel, EPA)
- Technical System Audits (TSAs) and Instrument Performance Audits (IPAs) of the National Air Toxics Trends Stations (NATTS) and Supporting Laboratories (S. Stetzer Biddle, Battelle)
- Interlaboratory Comparison of Ambient Air Samples (C. Pearson, CARB)
- Developing Criteria for Equivalency Status for Continuous PM2.5 Samplers (B. Coutant, Battelle)

**1:00 – 2:30 WEDNESDAY, APRIL 13TH**

**Environmental Laboratory Quality (ELQ)** (Salons A-C) *Chair: L. Doucet, EPA*

- Environmental Laboratory Quality Systems: Data Integrity Model and Systematic Procedures (R. DiRienzo, DataChem Laboratories, Inc.)
- The Interrelationship of Proficiency Testing, Interlaboratory Statistics and Lab QA Programs (T. Coyner, Analytical Products Group, Inc.)
- EPA FIFRA Laboratory Challenges and Solutions to Building a Quality System in Compliance with International Laboratory Quality Standard ISO 17025 (A. Ferdig, Mich. Dept. of Agriculture)

**Performance—Quality Systems Implementation (P-QSI)** (Salon D) *Chair: A. Belle, EPA*

- Implementing and Assessing Quality Systems for State, Tribal, and Local Agencies (K. Bolger, D. Johnson, L. Blume, EPA)

**1:00 – 2:30 WEDNESDAY, APRIL 13TH (continued)**

**Quality Initiatives in the EPA Office of Environmental Information (QI-OEI)** (Salons F-H) *Chair: J. Worthington, EPA*

- Next Generation Data Quality Automation in EPA Data Marts (P. Magrogan, Lockheed)
- The Design and Implementation of a Quality System for IT Products and Services (J. Scalera, EPA)
- Data Quality is in the Eyes of the Users: EPA's Locational Data Improvement Efforts (P. Garvey, EPA)

**A Win-Win-Win Partnership for Solving Environmental Problems (W3PSEP)** (Salon E) *Co-Chairs: W. Hunt, Jr. and K. Weems, NCSU*

- Overview of Environmental Statistics Courses at NCSU (B. Hunt, NCSU Statistics Dept.)
- Overview of the Environmental Statistics Program at Spelman College (N. Shah, Spelman)
- Student presentations: H. Ferguson and C. Smith of Spelman College; C. Pitts, B. Stines and J. White of NCSU

**Ambient Air Session III** (Sierra 5&6) *Chair: M. Papp, EPA*

- Trace Gas Monitoring for Support of the National Air Monitoring Strategy (D. Mikel, EPA)
- Comparison of the Proposed Versus Current Approach to Estimate Precision and Bias for Gaseous Automated Methods for the Ambient Air Monitoring Program (L. Camalier, EPA)
- Introduction to the IMPROVE Program's New Interactive Web-based Data Validation Tools (L. DeBell, Colorado State University)
- The Role of QA in Determination of Effects of Shipping Procedures for PM2.5 Speciation Filters (D. Crumpler, EPA)

**3:00 – 4:30 WEDNESDAY, APRIL 13TH**

**Topics in Environmental Data Operations (TEDO)** (Salons A-C) *Chair: M. Kantz, EPA*

- Ethics in Environmental Operations: It's More Than Just Lab Data (A. Rosecrance, Laboratory Data Consultants, Inc.)
- QA/QC of a Project Involving Cooperative Agreements, IAGs, Agency Staff and Contracts to Conduct the Research (A. Batterman, EPA)
- Dealing with Fishy Data: A Look at Quality Management for the Great Lakes Fish Monitoring Program (E. Murphy, EPA)

**Quality System Development (QSD)** (Salon D) *Chair: A. Belle, EPA*

- Development of a QA Program for the State of California (B. van Buuren, Van Buuren Consulting, LLC)
- Integrating EPA Quality System Requirements with Program Office Needs for a Practical Approach to Assuring Adequate Data Quality to Support Decision Making (K. Boynton, EPA)
- Introducing Quality System Changes in Large Established Organizations (H. Ferguson, EPA)

**Auditor Competence (AC)** (Salons F-H) *Chair: K. Orr, EPA*

- Determining the Competence of Auditors (G. Johnson, EPA)

**To Detect or Not Detect—What Is the Problem? (TDND)** (Salon E) *Chair: J. Warren, EPA*

- A Bayesian Approach to Measurement Detection Limits (B. Venner)
- The Problem of Statistical Analysis with Nondetects Present (D. Helsel, USGS)
- Handling Nondetects Using Survival Anal.(D. Helsel, USGS)
- Assessing the Risk associated with Mercury: Using ReVA's Webtool to Compare Data, Assumptions and Models (E. Smith, EPA)

**Ambient Air Session IV** (Sierra 5&6) *Chair: M. Papp, EPA*

- Status and Changes in EPA Infrastructure for Bias Traceability to NIST (M. Shanis, EPA)
- Using the TTP Laboratory at Sites with Higher Sample Flow Demands (A. Teitz, EPA )

**5:00 – 6:00 PM WEDNESDAY, APRIL 13TH**

**EPA SAS Users Group Meeting** Contact: Ann Pitchford, EPA

**8:30 – 10:00 THURSDAY, APRIL 14$^{TH}$**

**Evaluating Environmental Data Quality (EEDQ)** *(Salons A-C) Chair: M. Kantz, EPA*

- QA Documentation to Support the Collection of Secondary Data (J. O'Donnell, Tetra Tech, Inc.)
- Staged Electronic Data Deliverable: Overview and Status (A. Mudambi, EPA)
- Automated Metadata Reports for Geo-Spatial Analyses (R. Booher, INDUS Corporation)

**Satellite Imagery QA (SI-QA)** (Salon D) *Chair: M. Cusanelli, EPA*

- Satellite Imagery QA Concerns (G. Brilis and R. Lunetta, EPA)

**Information Quality Perspectives (IQP)** (Salons F-H) *Chair: J. Worthington, EPA*

- A Body of Knowledge for Information and Data Quality (J. Worthington, L. Romero Cedeno, EPA)
- Information as an Environmental Technology – Approaching Quality from a Different Angle (K. Hull, Neptune and Co.)

**To Detect or Not Detect—What Is the Answer? (TDND)** (Salon E) *Chair: A. Pitchford, EPA, Co-Chair: W. Puckett, EPA*

- Using Small Area Analysis Statistics to Estimate Asthma Prevalence in Census Tracts from the National Health Interview Survey (T. Brody, EPA)
- Logistical Regression and QLIM Using SAS Software (J. Bander, SAS)
- Bayesian Estimation of the Mean in the Presence of Nondetects (A. Khago, University of Nevada)

**Ambient Air Workgroup Meeting** (Sierra 5&6) *Contact: Mike Papp, EPA*
NOTE: This is an all-day, closed meeting.

**10:30 – 12:00 THURSDAY, APRIL 14$^{TH}$**

**Environmental Data Quality (EDQ)** (Salons A-C) *Chair: V. Holloman, EPA*

- Assessing Environmental Data Using External Calibration Procedures (Y. Yang, CSC)
- Groundwater Well Design Affects Data Representativeness: A Case Study on Organotins (E. Popek, Weston Solutions)

**Information Quality and Policy Frameworks (IQPF)** (Salons F-H) *Chair: L. Doucet, EPA*

- Modeling Quality Management System Practices to an Organization's Performance Measures (J. Worthington, L. Romero Cedeño, EPA)
- Development of a QAPP for Agency's Portal (K. Orr, EPA)
- Discussion of Drivers and Emerging Issues, Including IT, That May Result in Revisions to EPA's Quality Order and Manual (R. Shafer, EPA)

**Office of Water; Current Initiatives (OW)** (Salon D) *Chair: D. Sims, EPA*

- Whole Effluent Toxicity--The Role of QA in Litigation (M. Kelly, EPA, H. McCarty, CSC)
- Review of Data from Method Validation Studies: Ensuring Results Are Useful Without Putting the Cart Before the Horse (W. Telliard, EPA, H. McCarty, CSC)
- Detection and Quantitation Concepts: Where Are We Now? (Telliard, Kelly, and McCarty)

**Sampling Inside, Outside, and Under (SIOU)** (Salon E) *Chair: J. Warren, EPA*

- VSP Software: Designs and Data Analyses for Sampling – Contaminated Buildings (B. Pulsipher, J. Wilson, Pacific Northwest National Laboratory , R. O. Gilbert)
- Incorporating Statistical Analysis for Site Assessment into a Geographic Information System (D. Reichhardt, MSE Technology Applications, Inc.)
- The OPP's Pesticide Data Program Environmental Indicator Project (P. Villanueva, EPA)

**1:00 – 2:30 THURSDAY, APRIL 14TH**
**Information Management** (Salons A-C) *Chair: C. Thoma, EPA*

- Achieve Information Management Objectives by Building and Implementing a Data Quality Strategy (F. Dravis, Firstlogic)

**UFP Implementation** (Salon D) *Chair: D. Sims, EPA*

- Implementing the Products of the Intergovernmental DQ Task Force: The UFP QAPP (R. Runyon, M. Carter, EPA)
- Measuring Performance: The UFP QAPP Manual (M. Carter, EPA, C. Rastatter, VERSAR)

**Quality Systems Guidance and Training Developments (QSG)** (Salons F-H) *Chair: M. Kantz, EPA*

- A Sampling and Analysis Plan Guidance for Wetlands Projects (D. Taylor, EPA )
- My Top Ten List of Important Things I Do as an EPA QA and Records Manager (T. Hughes, EPA)
- I'm Here---I'm Free----Use Me! Use Me!—Secondary Use of Data in Your Quality System (M. Kantz, EPA)

**Innovative Environmental Analyses (IEA)** (Salon E) *Chair: M. Conomos, EPA*

- Evaluation of Replication Methods between NHANES 1999-2000 and NHANES 2001-2002 (H. Allender, EPA)
- Assessment of the Relative Importance of the CrEAM Model's Metrics (A. Lubin, L. Lehrman, and M. White, EPA)
- Statistical Evaluation Plans for Compliance Monitoring Programs (R. Ellgas, Shaw Environmental, Inc.; J. Shaw, EMCON/OWT, Inc.)

# Quality Assurance Documentation to Support the Collection of Secondary Data

**Ellis, S., O'Donnell, J., Tetra Tech**

## Introduction

In recent years, due to increased fiscal tension within the federal, state, and local environmental agencies, and with advances in information management and accessibility, collection of new environmental data, characterizations, and investigations have become an exception rather than the rule in environmental decision making. Use of data and information, including model outputs, collected by others or for other purposes (also called "secondary data") requires a comprehensive and detailed documentation system to support not only the collection of the information, but to ensure a traceable and auditable document trail detailing the unbiased evaluation and selection criteria for their subsequent use. It is clearly more cost effective to make use of available information and data than to start from scratch when addressing a new concern or investigative focus.. Collection of primary data are often relegated to those data that fulfill highly focused gaps in existing data, or for use in calibration and validation of predictive models so widely used to simulate and emulate natural processes and conditions in the environment. While the collection, evaluation, and use of data collected by others for other specific purposes encompasses not only environmental data, this discussion addresses primarily secondary data, but the same basic concepts can be applied almost universally to all secondary information collection, evaluation, and use. Simply, the collection, evaluation and use of secondary data requires a consistent and uniform treatment of all available information and data sources to ensure that data are not selectively reported, that all information sources of acceptable quality and caliber are evaluated against similar criteria, and that the evaluation criteria be documented in such a way as to make the process transparent for the current application and to future investigators and data users.

## Identification of Need

In a recent contract kick-off meeting with EPA staff, a number of topics were discussed as barriers to efficient contractual, fiscal, technical, and quality performance for some contract work and task assignments. Agency personnel responsible for fiscal management, contracting, and quality assurance expressed concerns about quality assurance requirements in the current agency culture; technical performance and flexibility under increasing fiscal and human resource limitations; and heightened documentation requirements mandated by the Office of Management and Budget and subsequent EPA Information Quality Guidelines. The frank and open discussion also included the challenges faced by contractors to adhere to the requirements of current contract language and limited budget resources while continuing to advance work assignments of significant importance to the EPA and the protection of human health and the environment. Potential solutions were solicited to address some of the many concerns expressed by all of the parties to the day's discussions. This presentation includes discussion of some of the key concerns expressed and offers some basic solutions which may alleviate some of the key issues discussed with agency staff.

**Proposed solutions**

The proposal for a generic secondary data collection QAPP including a model process for data collection and evaluation of available resources directly addresses a number of concerns voiced by EPA staff during the course of the discussion. Further, as the proposed generic process would include development of additional information management tools to enhance information collection operations within and outside of the Agency (how and where to access available data resources; the type and quantity of available resources; the categorization and assessment of available data based on the source and key characteristics; and the enhanced searchc capabilities enabled through expanded keywords and metadata descriptions accompanying the available data), the proposal was expanded to introduce the basic concepts to EPA quality staff and to solicit further input into the model process and on-line tools for potential development and implementation.

Among the concerns raised that may directly be addressed through development and implementation of a comprehensive secondary data collection guideline and an accessible, expanded information management tool were:

1) the contractual requirement for QAPP approval prior to commencement of Work Assignment or Task Order activities;
2) the Agency and its Prime contractor's flexibility to respond to dynamic funding conditions within the agency and the strained resources available to address the Agency's key mission objectives and details;
3) the graded approach to quality systems implementation and requirements to ensure that resources and application are appropriate to the needs of the specific program;
4) the impact of the Information Quality Guidelines documentation requirements, and defensibility of data and technical reports under IQG challenge; and
5) the basic quality culture that exists within the Agency, its regional offices, and the prime contractor communities with regard to quality assurance and quality control requirements and documentation.

On the development of a "generic" secondary data collection QAPP. The generic secondary data collection QAPP is more a process outline and documentation description than a quality assurance requirement. It is envisioned to include a logical process of discovery, definition, and exploitation of <u>all</u> available data resources; a graded, "screening" approach to establish a set of minimum standards for acceptance for the consideration of available data; and a set of uniform characteristics or data elements (depending on the data type) that can be scrutinized and graded in the consideration of data for further analysis. The generic process outline format enables project technical staff to proceed rapidly after award of a work or task assignment using an approved plan which includes sufficient documentation to monitor and verify performance objectives for data collection activities, while further development of data requirements and data quality objectives can be delayed based on the findings of the information collection operations. Development of QAPPs prior to or during the project planning phases is often difficult as project staff uncover valuable information that may directly or indirectly satisfy project information requirements. The process is developed with full consideration of information quality guidelines presented by the Office of Management and Budget (OMB) and EPA to ensure quality, objectivity, and integrity in the work products submitted to and disseminated by the federal government.

**Regulatory and other Guidance**

The Office of Management and Budget (OMB) through its Office of Information and Regulatory Affairs (OMB-OIRA) and in compliance with the directives in section 515 of the Treasury and General Government Appropriations Act for Fiscal Year 2001, published its *Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Information Disseminated by Federal Agencies.* In accordance with the requirements of Section 515, the guidelines

> ‘‘(1) apply to the sharing by Federal agencies of, and access to, information disseminated by Federal agencies; and
> ‘‘(2) require that each Federal agency to which the guidelines apply—
> ‘‘(A) issue guidelines ensuring and maximizing the quality, objectivity, utility, and integrity of information (including statistical information) disseminated by the agency, by not later than 1 year after the date of issuance of the guidelines under subsection (a);
> ‘‘(B) establish administrative mechanisms allowing affected persons to seek and obtain correction of information maintained and disseminated by the agency that does not comply with the guidelines issued under subsection (a); and
> ‘‘(C) report periodically to the Director—
> ‘‘(i) the number and nature of complaints received by the agency regarding the accuracy of information disseminated by the agency and;
> ‘‘(ii) how such complaints were handled by the agency.’’

In accordance with the OMB guidelines, EPA has issued its own *Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Information Disseminated by the Environmental Protection Agency EPA/260R-02-008.* These guidelines were further described as the Information Quality Guidelines (IQG), and they add specific details to the OMB’s general guidance for EPA’s procedures and processes, given the wide variety of information formats and conveyances for dissemination. EPA’s IQG is assembled in a logical format describing each of the critical components and systems for preservation of its information quality and integrity mission and policies.

Critical to both OMB and EPA’s guidelines are the assessment of transparency, clarity, consistency, and reasonableness (TCRR) characteristics of the information to be disseminated, as well as the information collection, evaluation, and use processes. Just as the TCCR characteristics define the quality of available data and information resources, they also must be engendered in a comprehensive assessment of numerous resources for consideration in specific data applications. By retaining the TCCR in the collection and evaluation processes and documenting the results of literature and data mining operations and evaluations, the data user is discouraged from selecting those sources which may be more supportive of a specific outcome or decision in their final use.  Transparency of the process is critical not only to ensure an unbiased process, but to ensure an auditable documentation trail for future investigators and data users.  In the opinion of the authors, the IQG, while providing the guidelines that support an unbiased approach to information collection and assessment, and mandating a documented approach to information collection and quality assessment, also suggests the opportunity for development of a more comprehensive tool that may promote and perpetuate a long-term benefit to EPA and the industry as a whole.

EPA Order 5360.1  CHG2- *Policy and Program Requirements for the Mandatory Agency-wide Quality System.*  Is the current underlying requirement for all EPA quality systems.  It mandates development of quality management plans for each of the agency's divisional and regional offices, as well as those of contractors providing contract support to the agency.

EPA Order 5360 A1 – *EPA Quality Manual for Environmental Programs,*  May 2000, EPA Office of Environmental Information Quality Staff.  describes the policy and requirements for compliance with the mandatory Agency-wide quality system (defined in Order 5360 CHG2).  Its guidance is more detailed and specific to the successful implementation of the general quality system described in Order 5360.  It is based on the national consensus standard authorized by the American National Standards Institute (ANSI) and published by the American Society for Quality (ASQ) in *ANSI/ASQ E4-1994 Specifications and Guidelines for Enviornmental Data Colleciton and Environmental Technology Programs*. The standard has been adapted as the basis for all agency and contractor quality systems for collection, evaluation, and use of environmental data and for the design, construction of environmental technologies.  ANSI/ASQ-E4 has since been updated to reflect more recent concerns of quality professionals in the US (ANSI/ASQ-E4-2004), however these revised, or, more specifically, more detailed guidelines have not yet been incorporated into the agency guidance or the agency-wide quality system.  Major changes reflected in the 2004 guidance included a more specific description of the independence of the QA function, senior management roles and responsibilities in the quality system implementation, and the need for a suitably flexible "graded approach" to quality system implementation to ensure that the quality system requirements are reflective of the variable importance of the specific environmental programs over which it is being implemented.  The graded approach is a practical response to the natural tension between operational and QA functions under increasing resource tensions experienced in all levels of the agency and in its contractor execution teams.  It ensures that the quality system retains the flexibility to respond to changing human and fiscal resource limitations which impact the various data collection or technology programs undertaken by the agency.  While these requirements are implicit in the 1994 guidance, they are described in greater detail in the 2004 revision of the standard.

In the scope of 5360 A1 the agency acknowledges that "environmental data are critical inputs to decisions involving the protection of the public and the environment from the adverse effects of pollutants from natural and man-made sources."  It further acknowledges the decisions often require the design, construction and operation of environmental technologies to protect human health and the environment from the deleterious effects of man-made and natural environmental pollutants.

## The Process Outline

The initial process is envisioned to include minimally a checklist and tracking sheet that would include some general, common-knowledge data resources, as well as those developed over multiple years of secondary data collection programs. The tracking sheet serves as documentation of the record search and the number and types of data that were discovered within a resource category, while the checklists are source-specific and detail the characteristics of the individual information sources, data sets, or technical papers. For specialized studies or data requirements, the tracking sheet may only include notation as to the availability or unavailability of applicable data sets, rather than the number and type encountered, or may continue to explore less well-known data resources, or evolve to a blank referral document where anecdotal information and references are identified through interviews and discussions with experts within a defined discipline. The tracking sheet documents fully the data collection process by resource type, and provides a coarse accounting of available data sources and available data sets, while the checklists add details as to source, type, key data and information presented, format, and other key characteristics to assist in the determerination of usefulness and applicability.

Given the wide variety of tasks and work assignments that require secondary data collections, the simple availability of data sets may determine the level of screening that can be applied (e.g., if no primary data collection funds are available, and limited secondary sources are identified, it may be necessary to use all of the sources; conversely, if copious amounts of data are available, it may be necessary to develop detailed screening and assessment protocols to ensure appropriateness of the data used in further analysis). Yes, there is such a thing as too much data.

By documenting the collection process, Agency Work Assignment Managers and Contractor Work Assignment Leaders are afforded the necessary tools to monitor the data collection process, and identify the key resources and challenges to a meaningful data collection. It is envisioned that eventually, these checklists will be able to be translated to electronic formats to be used without active research, and that they will produce valid data catalogues that can be indexed for future data collections and alleviate the need for significant human resource allocations to the information collection process.

In consideration of recent projects requiring reliance on secondary information and data, the authors arrived at a few basic, but inescapable conclusions in implementing the quality system requirements into the information collection and evaluation processes. While, in broad terms, it is easy to describe the process of information collection and evaluation, much can be taken for granted in the application of best professional judgement that may detract from the unbiased evaluation and selection of information resources for a specific application. Further, with consideration of EPA’s Information Quality Guidelines (IQG) policy and accompanying documentation requirements, consideration was given to the development of a generic secondary data collection Quality Assurance Project Plan (QAPP). The exploration of the documentation requirements to support the collection, evaluation, and use of secondary data revealed clear indication that a meaningful QAPP for secondary data collection lies within the associated checklists and documentation used in the collection and assessment of potential information resources.

To address some of the specific judgments which may otherwise remain under- or undocumented, the authors concluded that judicious use of a comprehensive checklist or collection of checklists will suitably document and support all of the decisions made during the collection and evaluation processes. Under either a comprehensive checklist or an adequately associated and cataloged series of checklists, much of the initial "screening" of information remains the same regardless of the information use. Information regarding the original sponsor, purpose, type, source, and format of information remain common to all information collection operations. Further, with regard to data collections, information detailing the name of the specific data collection program; whether the data quality objectives (DQOs) were prepared under a comprehensive planning process and are apparent (regardless of whether they are similar to or applicable under the current use requirements); whether data were collected under a quality assurance project plan (QAPP); whether data quality indicators (DQIs) are apparent (again, regardless of whether they are similar to or applicable under the current use requirements); whether field measurements and sampling operations were performed according to accepted practices and procedures are documented in accessible standard operating procedures (SOPs) and field sampling logs; whether laboratory measurements were conducted under approved methods, and whether they were made by laboratories certified in the use of such methods; whether method and sample QC data and calibration data are accessible; whether laboratory method and quantitation limits are apparent and verifiable; and whether data were subjected to third party validation, and what protocols were used to validate the results. While none of these questions may specifically qualify or disqualify data for a specific use on their own merit, they provide valuable descriptions of a data set that can be recorded and stored as part of the collection process and form a beneficial tool for future data collection efforts. (e.g., a high quality surface water data set may be lacking in geographical specificity required to model all point and nonpoint source loads in a given stream, but it may still provide a reasonable assessment of the water quality of the stream). A comprehensive checklist provides documentation of not only the strengths, but the deficiencies of the data set that will assist in its future consideration for other potential uses.

A limited number of metrics were considered to assist in an objective assessment of data quality and a scoring or ranking system discussed to first screen data for acceptance during the collection process, and, finally, to evaluate data and information for inclusion in further analyses and work products. The scoring or ranking of data may differ depending on the number of available sources, or of the intended use, however the scoring may be retained separate from the basic checklist responses and document description. A specific agency, grantee, or contractor organization, may develop internal policies regarding some of the scoring parameters, but retain the flexibility in the use characteristics, while others may modify the scoring for each specific data requirement as part of its collection process. Assignment of scoring or ranking values to data may well depend on their use or on the number of potential resources identified. While a given metric may disqualify a resource from use for one application, limited resources may dictate the need for use of all resources in another, hence the project planning and data quality objectives processes may determine the specific requirements for scoring or ranking available resources. The primary benefit of a ranking system is within the graded approach, whereby a change in funding or other condition may require a rapid reassessment of the selection criteria (i.e. downgrading or upgrading a composite score or rank).

The graded final data usability assessment is the point at which disparity becomes apparent in most secondary data collections, as the amount of available data are assessed with consideration with the quality and quantity requirements for a given task. However, certain key discriminators are common in all data assessments, even if they serve only to define their limitations. This process again, follows a basic logical progression of investigation and verification of key data elements and characteristics depending on the type of data required for the task. While a uniform grading system may remove the subjectivity of a comprehensive assessment, it may be inappropriate in order to ensure preserve the availability of the data for multiple uses, and, as such may require multiple assessments or scoring evaluations. It is therefore envisioned that the checklists developed for scoring and assessment of available data sets would include primarily the characteristics of a comprehensive data set, and their key discriminators to be assessed rather than guidance on how they are to be evaluated. The checklist would, again, include a series of narrative headings for descriptions of highly unique discriminators that may be applied to extremely high-profile or contentious data collections. Again this ensures the flexibility of the data user to incorporate all data where necessary, or to refine and develop highly descriptive criteria where requirements dictate. With clear identification of the data elements assessed or available for assessment, these checklists provide a catalog for consideration and assessment in future secondary data collection requirements. The scoring and assessment of each data set can be readily documented on the checklists with minor modification, thereby providing a fully defensible documentation of the elements evaluated and the scoring applied to each.

The secondary data collections are documented through the tracking sheet and checklists or a series of checklists identifying routine and non-routine data resources, including resources identified through interviews and anecdotal referrals. The individual resource checklists include the variety of key information categories and the data elements and characteristics that are generally considered minimum to each, the key characteristics required for secondary data assessment and analysis, and narrative discussion of key discriminators that may define one set of data as superior to another where similar information are available. It affords the contractor and the agency the documentation of the collection and evaluation processes, incorporating objectivity within a unique application, rather than attempting to grade data sets out of hand with only consideration of the data and its source. These documentation developed throughout the logical data collection operations ultimately provide the full documentation of data sources evaluated, those which were included in the collections, the elements evaluated for further analysis, and those that may require reconsideration through the iterative data analysis. Regardless of the ranking or scoring system which may or may not be applied, the “screening assessment” includes population of the basic data descriptions, while further scrutiny is more focused on the current study requirements, and may become more subjective. (i.e., different data collection purposes may not be suitable for all uses). In a comprehensive checklist format, a great deal of the information about a specific data set or candidate information source may be objective binary yes/no responses. It is the authors' opinion that these information could readily be captured and cataloged, perhaps by way of an on-line tool to enhance the current web inventory or to supplement other search engines or web tools which may be already available within the agency and contractor community. Further, by developing a web-accessible system, the requirement to populate the database could be incorporated into all work assignments that include secondary data collection, thereby rapidly creating a more robust database of searchable data characteristics for future data mining needs, and reducing the information collection burden of future work assignments.

**Example 'screening' checklist questions**

Was data/information provided by EPA?
Were data/information provided by another federal agency?
Were data/information provided by a state/territorial/tribal environmental agency?
Was data/information used in development of primary regulation?
Has data/information been published by EPA?
Were data/information subject to formal EPA peer review?
How were the data/information accessed?
What format are the data in? (electronic, hardcopy, database, text, etc.)
What type of information or data are presented? (geographical, physical, chemical, biological data or opinion, etc.)
What was the purpose of the data collection or information?
Were information collected under a QAPP?
Are DQO's presented or apparent?
Are any DQI's presented or apparent?

**Recommendations**

The requirements of the OMB and EPA's information quality guidelines ensure a transparent data collection and evaluation process for industry to compile, assess, and make use of data and information collected by others or for other purposes as part of the government's requirements for dissemination of information. These same guidelines suggest the application of a uniform process for all information collection requirements and documentation of their associated evaluation for use. The process outline proposed in the form of a generic QAPP includes thorough documentation through judicious use of comprehensive process tracking and information checklists and provides an excellent foundation from which additional tools can be developed to assist the agency, grantees, and contractors in their information and data mining operations. Members of EPA's quality staff, state environmental agencies and the grantee and contractor communities are invited to collaborate in the development of comprehensive tracking systems and checklists for the monitoring of data and information collection operations. Development of these tools and incorporation of their use into routine work and task assignments affords the dedication of important fiscal and human resources toward their primary environmental project objectives.

## REFERENCES

Office of Management and Budget. 2002.. *Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Information Disseminated by Federal Agencies.* (67 FR 8452)

U.S. Environmental Protection Agency. 2002. EPA/260R-02-008. *Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Information Disseminated by the Environmental Protection Agency.* EPA/260R-02-008. EPA Office of Environmental Information, Washington, DC.

U.S. Environmental Protection Agency. 1999. EPA Order 5360.1 CHG2, *Policy and Program Requirements for the Mandatory Agency-wide Quality System*. EPA Office of Environmental Information, Washington, DC.

U.S. Environmental Protection Agency, 2000 EPA Order 5360 A1, *EPA Quality Manual for Environmental Programs*. EPA Office of Environmental Information, Washington, DC. Washington D.C.

American National Standards Institute. 1994. ANSI/ASQ E4-1994, *Specifications and Guidelines for Environmental Data Collection and Environmental Technology Programs*. 1994. American Society for Quality, Milwaukee, WI.

American National Standards Institute. 2004. ANSI/ASQ E4-2004, *Specifications and Guidelines for Environmental Data Collection and Environmental Technology Programs*. 2004. American Society for Quality, Milwaukee, WI.

# Staged Electronic Data Deliverable (SEDD) – Overview and Status

Anand R. Mudambi

US EPA
Office of Superfund Remediation and Technology Innovation,
1200 Pennsylvania Ave NW, Mail Code 5102G,
Washington, DC 20460

E-mail Address: **mudambi.anand@epa.gov**

## The need for Open Data Standards and Uniform Formats for Electronic Transmission of Environmental Data

The Federal Agencies including the US Environmental Protection Agency, US Department of Defense, and US Department of Energy collect large amounts of data to make environmental decisions like extent of site contamination, cleanup remedies, site remediation end points.  In order for this information collection to be efficient and cost effective, data collected in electronic formats are the preferred option due the ease of transmission, receipt, evaluation, storage, and retrieval.

Many of the benefits of electronic data are nullified if they are transmitted in proprietary formats since data cannot always be exchanged between various groups.  This data exchange is very important since many environmental decisions are taken based on data that is collected by one Federal entity and then reviewed by many others.

There is thus a strong Federal need to receive electronic data used for environmental decisions in:

a. a non-proprietary open data standard formats like HTML (Hyper Text Markup Language) and XML (eXtensbile Markup Language)

b. a uniform electronic format especially for environmental data that needs to be exchanged.

## Problems for Environmental Laboratories

In today's information age, the environmental laboratories have to report data to most of their clients (including Federal Agencies) in an electronic format.  These formats can range from simple electronic spreadsheets to complex ones like the US Air Force Installation Restoration Program Management System (IRPMS) and US EPA's Agency Standard Format (ASF) electronic deliverables.  The information conveyed by the electronic data deliverables (EDDs) also varies widely depending on client needs.  Laboratories routinely have to support a myriad of reporting formats (in some cases over 100 electronic formats), which increases their operating costs.  These formats are also constantly changing as client requirements and methods change adding even more costs to an already burdened industry.

The different types of electronic deliverables also pose problems for the laboratory clients (including Federal and State Agencies), which have to create different electronic tools to evaluate the EDDs. It is very expensive to develop and maintain these tools, which can then be used only for the EDDs for which they were created. With decreasing budgets, it becomes even more vital for Federal Agencies to share electronic information and this becomes difficult (if not impossible) if the EDDs are incompatible with the different Agency tools.

## What is the SEDD Specification?

SEDD stands for Staged Electronic Data Deliverable. The SEDD Specification provides a common structure and data element dictionary to report a wide variety of data (chemical, radio chemical, biological, etc.) to multiple customers. The SEDD Specification allows for reporting of analytical data in multiple formats ranging from simple sample concentrations all the way to a CLP type data package and beyond. The SEDD Specification views reporting of analytical data in the same manner as the laboratory produces it - i.e., it is based on the way data is generated in the laboratory for the analysis of a sample. The SEDD Specification is thus designed for reporting the laboratory analytical data and results along with enough information to be able to connect these laboratory results to the site specific sample information taken during the collection of the sample in the field.

The SEDD Specification consists of the following documents:

- An Overview Guide which gives the specifications and structure of creating a SEDD file. Creating a SEDD file requires the use of XML technology and EDDs created using the SEDD Specification are transmitted as XML documents. XML is an open Data Standard and stands for eXtensible Markup Language. It provides a common approach for transmitting information over the Web. This language is a Final Standard recommended by the World Wide Web Consortium (W3C).

- A Data Element Dictionary that gives the SEDD data elements, their corresponding definitions and allowed valid values.

The latest versions of these documents are available at the following website:

**www.epa.gov/superfund/programs/clp/sedd.htm**

Both the Overview Guide and Data Element Dictionary are agency and program neutral - i.e., they do not contain biases or requirements for any particular agency or program.

## Common SEDD Misconceptions

The SEDD XML document is not a database used for generating electronic reporting formats like Laboratory Information Management Systems (LIMS) or used for receiving, storing and retrieving environmental data like the US EPA’s STORET (STOrage and RETrieval) and

SDWIS (Safe Drinking Water Information System).

## What is a SEDD File?

1. A SEDD electronic file is a hierarchal file created by a laboratory from their information management system (a single or multiple databases) and is based on the SEDD Specification.

2. A SEDD file contains information regarding the chemical analysis of sample(s).

3. A SEDD file is a XML document (with an .xml extension).

## SEDD Stages

From the SEDD Specification four (4) specific EDD formats (stages) have been created. These individual formats are unique in that each stage directly builds on the previous stage allowing the user to specify the level of detail as needed for a given program or project. These SEDD stages lay out the reporting requirements for a SEDD electronic data deliverable that is agency and program neutral. Thus it can be used for data exchange between agencies and programs.

Stage 1 (Figure 1) only uses a small part of the overall SEDD structure and contains a minimum number of data elements to transmit results only data.

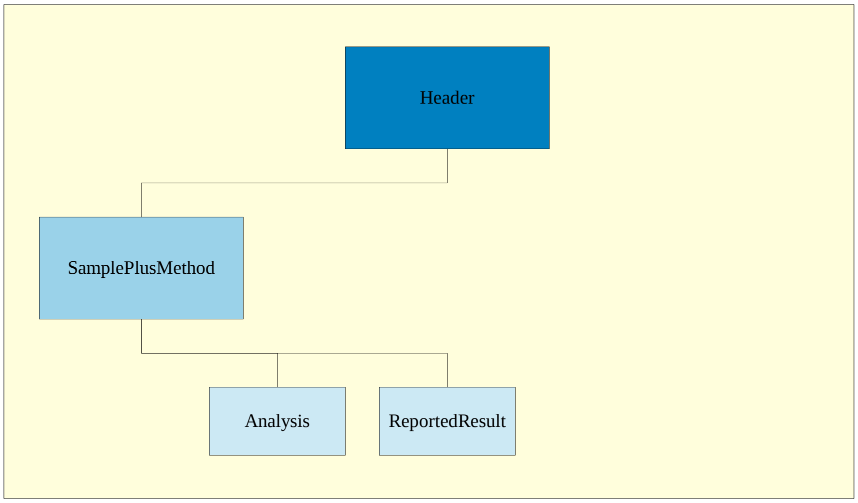


Figure 1. Structure of SEDD Stage 1

Stage 2 contains all of the Stage 1 structure and data elements but adds additional structural and data elements to report method quality control (Stage 2a – see Figure 2) and instrument quality control (Stage 2b – see Figure 3) information.

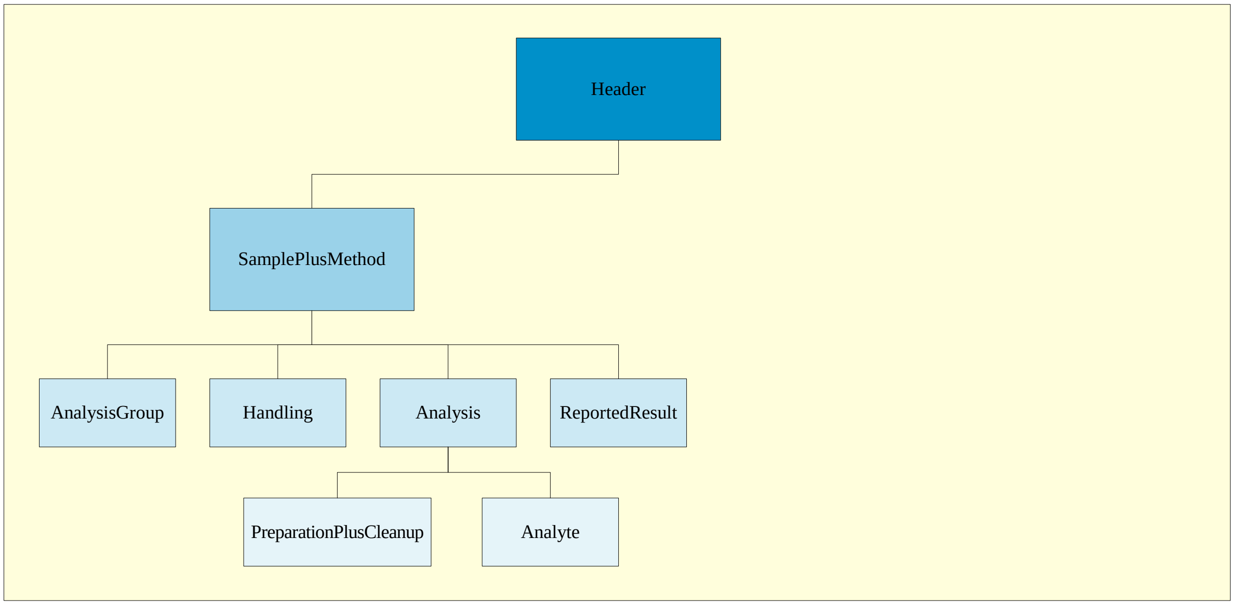


Figure 2. Structure of SEDD Stage 2

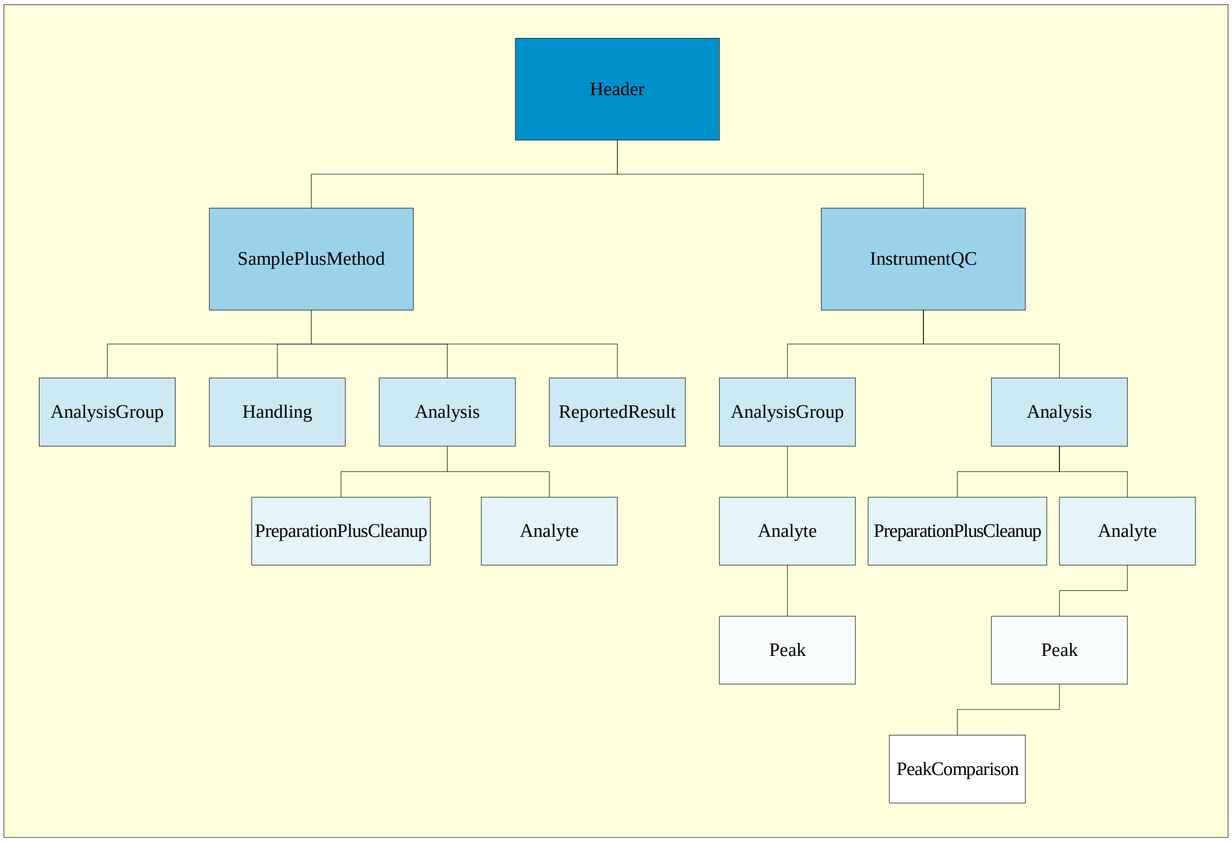
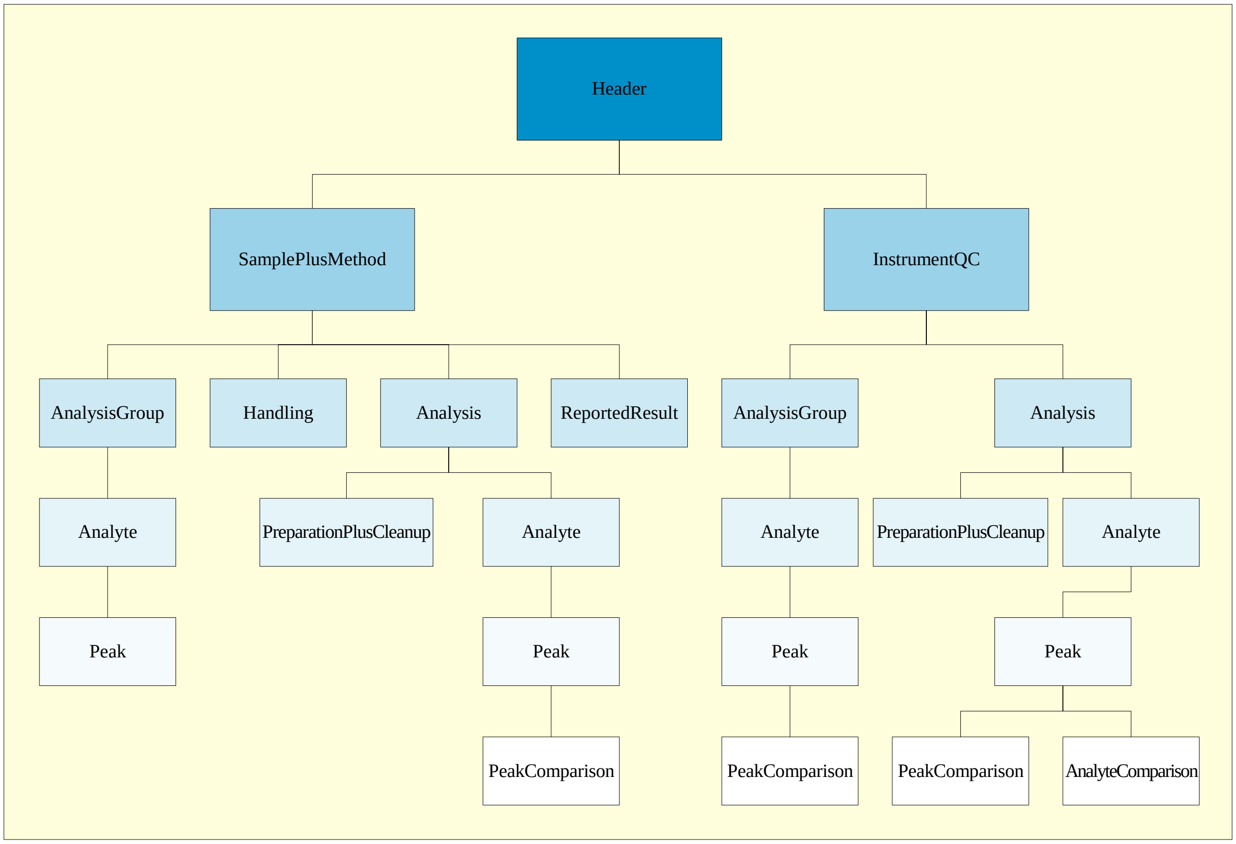


Figure 3.  Structure of SEDD Stage 2b

Stage 3 (Figure 4) contains all of the Stage 2 structure and data elements but adds additional structural and data elements to allow for the independent recalculation of the reported results (e.g., as required by CLP).



A fifth format (Stage 4) is now under development that would build on Stage 3 and allow for the reporting of all raw instrument data files.

## Example XML Files

An example XML file for reporting the final results for benzene as analyzed using a typical GC/MS method would look as follows:

```
<Results>
        <ClientAnalyteID>Benzene</ClientAnalyteID>
        <CASNumber>71-43-2</CASNumber>
        <Result>24.2</Result>
        <ResultUnits>ug/L</ResultUnits>
</Results>
```

The EDD consists of a series of data elements that are nested within the various structural elements (nodes). All data elements within the SEDD specification use a tagged format and contain the actual analytical information. Each data element uses real words rather than codes such that they are readily understandable by others. For example one of the data elements used in SEDD is ClientAnalyteID, which stands for “A client defined code for an analyte” in the SEDD Data Element Dictionary (i.e., what the client calls the analyte). This data element would contain the name of an analyte (e.g., Benzene) as recognized by the client.

An example XML file for reporting the preparation information for the separatory funnel extraction of a liquid sample that will be analyzed using a typical GC/MS method would look as follows:

```
<Preparation PlusCleanup>
        <ClientMethodID>3510C</ClientMethodID>
        <PreparedDate>03/06/2003 08:00</PreparedDate>
        <AliquotAmount>1.00</AliquotAmount>
        <AliquotAmountUnits>L</AliquotAmountUnits>
        <FinalAmount>1.0</FinalAmount>
        <FinalAmountUnits>mL</FinalAmountUnits>
        <PreparationBatch>WG12114-03/06/2003-1</PreparationBatch>
</PreparationPlusCleanup>
```

Another feature of the SEDD specification is that it allows for the unambiguous linking of all QC samples to the regular samples. For this example, all samples (both QC and regular) that contain the value 'WG12114-03/06/2003-1' as the value for the PreparationBatch data element would be linked together.

Since XML technology is being used, files generated using the SEDD specification can be readily viewed/edited using third party software products. An example of such a free viewer/editor is XML Notepad as written by Microsoft.

## Advantages of Using SEDD (Includes Cost Savings)

There are many advantages for both laboratories and data requesters like Federal Agencies when SEDD is used as the data transmission format. SEDD reduces the number of EDDs laboratories currently have to support since SEDD can meet multiple agency requirements. For Federal Agencies, a common EDD allows for development of common automated data review tools to check the EDDs. These tools can be then shared across agencies.

There are also significant cost savings when SEDD files are sent and reviewed by electronic software. Preliminary results show a 30 to 50% cost savings when SEDD files are electronically reviewed when compared to similar manual reviews.

## SEDD Inter Agency Efforts

Offices from the US EPA, US Army Corps of Engineers, US Air Force, US Navy, US Department of Energy, and others are cooperating to review and/or pilot this Specification for delivery of environmental chemical and radio chemical data. Face-to-face meetings, conference calls and video conferences are being held on as needed basis to ensure that the SEDD Specification can meet program specific needs (while remaining generic enough for data exchange between the agencies).

Information regarding these inter agency efforts can be obtained on both the previously identified EPA website and the following USACE website:

www.environmental.usace.army.mil/info/technical/chem/chemedd/chemedd.html

The web sites also contain several specific example SEDD XML documents for various analytical testing methods at various stages (stages 2a, 2b, and 3). These examples can be used by laboratories to see how the SEDD XML documents are structurally assembled and linked. In addition, they can be used to test new automated review tools.

For each example SEDD XML document, a Document Type Definition (DTD) file and a specific Instruction file (as a Spreadsheet) are posted. The DTD lists the specific parts of the SEDD structure and the specific data elements to be used for that EDD. The Instruction files show what specific data elements are to be used for each sample type likely to be encountered when the referenced method is used.

## SEDD Implementation Status

Offices from the US EPA and US Army Corps of Engineers have been conducting pilot projects with laboratories for Electronic Data Deliverables based on the SEDD Specification since May 2002.

USEPA now requires the delivery of SEDD Stage 3 files for the Contract Laboratory Program. The delivery of a Stage 3 EDD will allow for full independent recalculation of the reported results from raw data. USEPA Superfund Technical Assessment and Response Team (START) contracts will also require SEDD files starting from April 2005.

The US Army Corps of Engineers (USACE) now requires the delivery of SEDD files for the FUDS (Formerly Used Defense Sites) Program. USACE contracts are being modified to meet this requirement. Several Projects have been competed which successfully parsed these EDDs into various data assessment systems and databases .

In addition, private sector companies are already creating software to receive and review SEDD files.

## Expanded Applications of SEDD

Currently the SEDD Specification has been developed to report analytical chemistry testing data. This data can be linked back to the field information (like sample location) through a limited number of SEDD data elements (e.g., ClientSampleID). The SEDD Specification can be expanded to include field generated information (e.g., sampling specific information including location and temporal data) and project information (e.g., Project Name, location, duration of operation).

In addition, the SEDD Specification already has data elements that apply to other types of environmental testing and reporting like radiochemistry, biological, and other methods. It can also be applied for reporting data from other areas besides environmental testing like the pharmaceutical and chemical manufacturing industries.

## Contact Information for SEDD

Please contact Anand R. Mudambi (US EPA) or Joseph Solsky (US Army Corps of Engineers) for more information regarding the SEDD Specification, SEDD Pilot Projects, SEDD Interagency Efforts or development of tools for evaluating and processing EDDs based on the SEDD Specification.

# AskWATERS
# Enhanced Metadata for Geo-Spatial Analyses

Increasingly, the realm of geo-spatial analyses is moving away from the dedicated geographer to general staff. With the use of web-based mapping tools, both the availability and capability to perform geo-spatial analyses have been given to general staff who are not familiar with the limitations of the technology and the limitations of the data used in the analyses. Although metadata may be present to properly describe the analysis and data used, without it being presented in a readily available contextual format that is easily understood, general staff can easily overlook this vital information asset and make erroneous conclusions from their analyses. Currently, the Environmental Protection Agency (EPA) is using the results of geo-spatial analyses to determine progress on Government Performance and Results Act (GPRA) measures and to set priorities for Agency resources. Since these analyses are critical to the Agency, it is essential that all analyses be accompanied with the properly formatted metadata.

AskWATERS is a reporting tool being developed by the Office of Water (OW), whose primary purpose is to provide automate reports of water based GPRA measures and whose secondary purpose is to provide a repository for geo-spatial analyses that utilized the Waters Tracking and Environmental Results System (WATERS) database. WATERS, is the official EPA store of water based entities (Permit Compliance System facilities, impaired waters, drinking water intakes, STORET monitoring stations, etc.) that have been spatially indexed to the National Hydrography Database (NHD). NHD, maintained by USGS and EPA, is the spatial representation of the network of rivers, streams, lakes, and ponds in the United States. With the spatial indexing of these water based entities to the NHD, geo-spatial analyses can be performed to answer questions such as “What PCS facilities are co-located on an impaired water?” and “What drinking water intakes do not have uses and monitoring criteria that are appropriate to protect public health?”. A design goal of AskWATERS is to provide tailored, relevant metadata with each report so that the average user can make an informed decision about their use of the reported analysis. For each report, an automated metadata report is generated that describes the following:

- Description of the data source used, containing information on the primary source of the data, how it was collected, and a data dictionary of all data reported.
- Description of the analyses performed.
- Currency of the data used in the analysis and how this currency relates to the currency of data in the primary data source.
- Currency, completeness, and validation (or lack thereof) of the spatial data used in the analysis.
- Links to external metadata resources.

For this technical paper, the metadata issues for AskWATERS will be described in detailed and a complete description of how these issues are addressed to provide automated metadata reports will be shown. If an internet connection is available, the AskWATERS system will be demonstrated with live metadata reports.

# REMOTE SENSING:
# QUALITY ASSURANCE AND ERROR PROPAGATION

George M. Brilis,
Ross Lunetta,
John G. Lyon
Environmental Sciences Division
National Exposure Research Laboratory
Office of Research and Development
U.S. Environmental Protection Agency
P.O. Box 93478
Las Vegas, NV 89193-3478

For the
24th Annual National Conference on Managing Environmental Quality Systems
San Diego, CA
April 11-15, 2005

## Introduction

The US EPA has been recognized by many organizations, private and public, as having one of the most extensive and effective Quality System. Remote Sensing (RS) Quality Assurance/Quality Control (QA/QC) is umbrellaed under Geospatial Science. The EPA Quality System has been supportive of the US EPA Geospatial Quality Council (GQC) since its' grass-roots formation in 1999. The GQC has developed a number of Agency-wide guidance documents and training courses that can be found at the GQC Internet site http://www.epa.gov/nerlesd1/gqc/default.htm.

This quality perspective is not intended to address all quality issues, nor provide all possible solutions that may be raised. Few documents to date address RS QA/QC. The main intent of this perspective is to raise the level of awareness to quality issues in RS.

Typically the term "RS images" conjures thoughts of satellite images. In reality, many historical and low-altitude images are aerial photographs. In many cases, historical aerial photographs are converted to digital images. New aerial photographs use digital technology.

RS images are, to a high degree, digital in origin and format. When high visibility concerns are involved, such as litigation in the Regions and patents in the Office of Research and Development (ORD), the following issues, and more, invariably come forth:

1. Intellectual property (IP) issues - rights of copy, "fair use," etc
2. Admissibility of image enhancement - the need and extent of enhancement

3. Composite image issues - which image was used first, to what extent was another used, and which is then the “derived”image
4. QA/QC concerns - image authenticity, veracity, integrity, and more.

The above-stated issues will change and evolve with technology, law, use and misuse of RS images. Therefore, these issues should set the foundation of the implementation agenda. As Information Technology (IT) changes, theoretical and practical problems arise from the science/law interface. The science/law interaction and effect on environmental science and decision making is discussed in the EPA Scientific and Technological Achievement Award-winning paper *Quality Science in the Courtroom: A Comparison of EPA Data Quality and Peer Review Policies and Procedures to the Daubert Factors*, Brilis et al

**Room for Error**

The products of RS, satellite images and aerial photographs, are often used in conjunction with other products and processes. These RS image processing software programs may include, but are not limited to, “ENVI,” ERDAS,” or a more general term, “Image Analysis” software. After processing, the resultant image may be used in yet another software application - such as a Geographic Information System (GIS). It is clear then, that RS image processing provides many opportunities for the introduction of human and computer error and the degradation of the image quality and/or integrity. In addition, the door is also opened to science misconduct. Human or technological errors can be introduced at the planning, acquisition, storage, transfer, processing, output, interpretation, and use stages of the RS image lifecycle. The general lifecycle flow of RS data can be described as:

- Planning
- Data Acquisition
- Data Input
- Storage of Data
- Data Transformation (or Manipulation)
- Out of product(s)
- Use
- The system loops back to planning if and when appropriate

**Common Errors in Satellite and Remotely Sensed Images**

The GQC has through audit results and interview with scientists, found the following to be the most common errors made in RS.

<u>Data Collection</u>

- Inaccuracies in the photogrammetric methods used to draw maps and measure elevations
- Image interpretation introduces a degree of error in the classification and delineation of boundaries.
- Inaccuracies in the photogrammetric methods used to draw maps and measure elevations

- Image interpretation introduces a degree of error in the classification and delineation of boundaries.

Data Input

- Center of a digitizing table has higher positional accuracy than the edges.
- Curved boundaries are actually small straight lines. Natural boundaries do not exist as a sharp line. Smaller lines create larger data files

Data Storage

- Commonly used storage form in vector-based GIS is the 32-bit real number format. This provides room for about 7 significant numbers. UTM uses 7 significant numbers. A data base that contains info with levels of detail ranging from fractions of a meter to full UTM would require greater precision. To retain accuracy of this diverse data, more than seven significant numbers would be needed. Solution is to store data in 64-bit format. This increases the volume of data, which yields larger data files and subsequently greater cost. Keeping cost down means less significant figures and therefore reduced accuracy.
- Storage in the form of Raster-based data introduces even more error. Raster uses a pixel to represent a unit of terrain. If data are encoded using a pixel size of 10m x 10m, then even if the location point is known to a fraction of a meter, it can only be represented to the nearest 10m. This yields a loss of accuracy unless one simply decreases the pixel size (from 10m x 10m to 1m x 1m) for greater accuracy. But this increases the number of pixels and the resulting file size.
- Generally, file size increases by the square of the resolution. So, increasing the resolution 10 times from 10m pixels to 1m pixels increases the file size about 100 times.
- For both Vector and Raster, there is a direct cost to keep higher levels of precision as a result of the increased storage.
- Vector is better for storing high precision coordinates for discrete map elements.
- Raster is better for representing measurements that vary continuously over an area.

Data Manipulation (Transformation)

- RS images are often used as the foundation upon which other RS images or other data sets will be overlaid. As the number of overlays increases, the greater the opportunity for errors to arise and propagate.
- The same boundary may be drawn slightly different in two overlays (one using short lines, the other using long lines). This mismatch will create inaccuracies in the resultant image.

Data Output

- Error can be introduced by the printing device, such as color or tone differences, and the resolution capability of the printing device.

- Media – Paper shrinks and swells. On a small scale map, the millimeter changes can represent several meters at the ground resolution.

Use of Results

- RS images are often incorporated with other data via a GIS. The results may be misinterpreted, accuracy and scale levels ignored and inappropriate analyses accepted

**Conclusion**

The above-listed sources of error are by no means complete. One must always consider the impact of errors on a case-by-case basis. Especially in a research environment, one must highlight the intended use of the study where RS was used. The processing of RS data does not take a route of "one-size-fits-all." In each image, the RS or Geospatial Professional must exercise judgment based on their individual education and experience. It is critical that the RS or Geospatial Professional document these important judgment junctures in order to ensure that another scientist can, to a reasonable extent, reproduce their results.

Some suggestions for the QA Professional to keep in mind when evaluating the quality of an RS project and/or product are:

- What quality-impacting events are within control of the scientist (certainly an EPA scientist cannot tune the wavelength sensitivity of a satellite senor)?
- Did the scientist document critical judgment junctures?
- What was the intent of the study and are the results being applied as intended?

Again, documentation is critical for the scientist and to ensure reproducibility. Emphasis on documentation by the QA Professional is one way to ensure the longevity of usefulness of the resources applied to projects that involve the RS.

# A Body of Knowledge for Information and Data Quality

*Jeffrey Worthington - OEI Director of Quality, Office of Planning, Resources, and Outreach, Office of Environmental Information, US Environmental Protection Agency*

*Lorena Romero Cedeño – Program Analyst, Office of Planning, Resources, and Outreach, Office of Environmental Information, US Environmental Protection Agency*

## Abstract:

The quality profession has long recognized that the breadth of quality practices constitute a well established discipline with a knowledge base that includes components of management, assessment, planning, reliability, etc.  Likewise, information science constitutes a growling knowledge base including hardware, software, planning, etc. Now, there is recognition that application of quality principles to information is evolving into a unique body of knowledge.  Individuals and groups working in the area of information and data quality have outlined several models and approaches for applying quality principles to information.  This presentation reviews and summarizes existing resources and presents a general model for an *information and data quality body of knowledge* that can be used by managers, quality managers, and planners as they work to integrate information quality into the organization's management and quality systems.

# Information as an Environmental Technology – Approaching Quality from a Different Angle

*Kevin Hull, Senior Quality Assurance Specialist, Neptune and Company*

## Abstract:

For decades the primary emphasis of EPA's quality program has been on the collection of environmental data and associated data quality descriptors. More recently, EPA quality managers have also begun to focus on the quality of environmental technology applications – hence the January 2005 publication of *Guidance on Quality Assurance for Environmental Technology Design, Construction and Operation* (QA/G-11), a document that is based on the environmental technology section of the ANSI/ASQ E4-2004 national consensus standard.

It is possible to bridge the gap between these two categories of quality management by thinking of *information* – the usable product of data collection activities – as itself an environmental technology. Like other environmental technology applications, information operations follow a life cycle, comprising planning, design, construction/fabrication, operation, assessment, and acceptance. This presentation will highlight the applicable section of the E4 standard to illuminate how a different set of quality concepts and tools can improve the management and use of environmental information.

# Using Small Area Analysis to Estimate Asthma Prevalence in Census Tracts

Thomas M. Brody, Lawrence Lehrman, Paul Levy,
Alan Walts, Edward Delisio, Betsy Smith

## Background

Several EPA programs need to make decisions on how to best use their resources to effectively reduce the burden of asthma at local levels across the United States. One such program is the Office of Environmental Justice. Recently, OECA policymakers released an Environmental Justice targeting strategy that calls for information on health, compliance, environmental, and demographic data[1]. Asthma was one of the supplemental health outcomes suggested for the targeting approach.

OECA's Environmental Justice targeting strategy points to an American Lung Association (ALA) report on Morbidity and Mortality for their asthma indicators[2]. The report mentions the National Health Interview Survey (NHIS), Behavioral Risk Factor Surveillance System (BRFSS), National Hospital Ambulatory Medical Care Survey, and other surveys as sources for estimating the asthma burden at the national level. However, the ALA and others acknowledge that measurements of the number of persons with asthma at the community level are generally not available[3,4].

To this end, the ALA used synthetic estimation techniques to assess the asthma burden[5]. In general, the synthetic method creates an estimate of the population having a health characteristic in a small area by applying proportions of the population having the health characteristic in one or more demographic categories (age, sex, race, etc.) in a larger area to population figures for these demographic categories in the small area. These techniques were originally developed by the U.S. Bureau of the Census and the National Center for Health Statistics in the 1960s[6].

[1] US EPA, *Environmental Justice Smart Enforcement Targeting Strategy*, November 2004. EPA300R04003.
[2] American Lung Association. Trends in Asthma Morbidity and Mortality, April 2004. Accessed January 14, 2005. Retrieved January 14, 2005 from http://www.lungusa.org/atf/cf/{7A8D42C2-FCCA-4604-8ADE-7F5D5E762256}/ASTHMA1.PDF
[3] American Lung Association. Estimated Prevalence and Incidence of Lung Disease By Lung Association Territory, September 2004. Retrieved January 14, 2005 from http://www.lungusa.org/atf/cf/{7A8D42C2-FCCA-4604-8ADE-7F5D5E762256}/ESTPREV2004.PDF
[4] Brody, T.M. (2004, May). Using Small Area Analysis to Estimate Asthma Prevalence in Chicago Public Schools. Public Health GIS News and Information (58), pp. 10-13. Retrieved January 14, 2005 from http://www.cdc.gov/nchs/data/gis/cdcgis58.pdf
[5] American Lung Association Estimated Prevalence and Incidence of Lung Disease By Lung Association Territory, September 2004. Retrieved January 14, 2005 from http://www.lungusa.org/atf/cf/{7A8D42C2-FCCA-4604-8ADE-7F5D5E762256}/ESTPREV2004.PDF
[6] Levy, P.S. (1979, Feb). Small area estimation--synthetic and other procedures, 1968-1978, National Institute on Drug Abuse Research Monograph (24), pp 4-19.

Various health departments including the Centers for Disease Control still use synthetic estimation techniques today for predictive capability[7].

The ALA used a very simple algorithm in their estimation method. Local area prevalence of pediatric asthma was estimated by multiplying the national prevalence rate from the 2002 NHIS for children under 18 by the under 18 county-level resident populations for the same year. The result was an estimate of the number of children affected by the disease in each county. Unfortunately, the proportionate burden could not be shown as the rate is inherently the same for all counties.

If OECA needs to compare the proportionate burden between areas, it is necessary to show the naturally fluctuating rates throughout the communities. Rates will fluctuate if more factors are taken into account. By including multiple significant social factors in the synthetic model, the rates for each area will fluctuate as these associated variables fluctuate throughout the Census Tracts. The result is a baseline estimate of what we might expect the asthma rate to be in a community before environmental factors are taken into account. These estimates would effectively answer the question of where we might want to reduce environmental factors because of the endogenously disproportionate asthmatic population in the community.

## Methods

Conceptually, the synthetic estimator can be written as

$$X_t = \sum_{\alpha=1}^{k} P_{t\alpha} X_\alpha$$

where

$X_t$ = the mean rate of characteristic $X$ for Census Tract $t$.

$P_{t\alpha}$ = the proportion of the population in tract $t$ who are members of population cell $\alpha$ (alpha). An alpha cell in this analysis is a demographically bounded class of age, sex, race, and income categories.

$X_\alpha$ = the mean rate of characteristic $X$ for persons in cell $\alpha$.

The underlying rationale for this model is that the distribution of a health characteristic does not vary among populations of the Census Tract except to the extent that the associated social factors among Census Tracts vary in demographic composition[8]. Social factors such as age, race/ethnicity, sex, and income all have been shown to be associated with asthma[9]. For this work, demographic proportions of the 2000 NHIS were synthesized with similar demographic

[7] CDC, Forecasted State-Specific Estimates of Self-Reported Asthma Prevalence -- United States, 1998, Morbidity and Mortality Weekly Report, Retrieved January 14, 2005 from http://www.cdc.gov/mmwr/preview/mmwrhtml/00055803.htm

[8] Levy, P.S., D.K. French (1977, Oct). Synthetic Estimation of State Health Characteristics Based on the National Health Interview Survey Data Evaluation and Methods Research (2)75. Hyattsville, MD: U.S. Department of Health, Education and Welfare. National Canter For Health Statistics.

[9] Ritchie, I.M, and R.G. Lehnen. The Indianapolis Asthma Study: Health Effects of Ozone and Other Environmental Measures on Children in the Indianapolis Metropolitan Area, Report submitted to the Indiana Department of Environmental Management, December 2001. Retrieved January 28, 2005 from http://www.in.gov/idem/planning/publications/indplsasthmastudy.pdf

categories within the year 2000 Census Tracts to estimate the burden of asthma throughout the Census Tracts of the 48 contiguous States.

In the NHIS process, families are randomly selected to answer the questionnaire in a personal household interview[10]. One sample adult and one sample child, if any, are selected from each family. Information on each is collected for the Sample Adult Person Section (SAP) and Sample Child Person Section (SCP) of the NHIS.

The 2000 SAP asked 32,374 adults the following questions.

1. Has a doctor or other health professional EVER told you that you had asthma? 3,052, or 9.4%, responded "yes."
2. During the past 12 months, have you had an episode of asthma or an asthma attack? 1,179, or 3.6% responded "yes."
3. During the past 12 months, did you have to visit an emergency room or urgent care center because of asthma? 342, or 1.1% responded "yes."

The SCP presented similar questions to adults concerning children in their household. The SCP accounted for 13,376 children. 1,630 or 12.2%, reported that a doctor or health professional had told them that the child in their household had asthma. 740, or 5.5%, reported that the child had an episode of asthma or an asthma attack during the last 12 months. 259, or 1.9%, reported the child visited the emergency room or urgent care center because of asthma.

Several other questions are asked of each sample adult and child in the NHIS including their age, race, sex, and family income. Using these additional data, the total population and the total asthmatic population were gathered for each combination of three age groups ("Under 18," "18 to 64," "65 and Over"), two gender groups ("Male," "Female"), four race/ethnicity groups ("White," "Black," "Hispanic," "Other"), and two income groups ("At or Below Poverty Level," "Above Poverty Level"). To provide stability in the estimates, each of the 3x2x4x2 = 48 combinations, or alpha cells, was assessed to ensure it had greater than 30 persons in total population. Attempts were made to further segment Race/Ethnicity into groups such as Native Alaskan, Hawaiian Islander, Native American, and Puerto Rican. Unfortunately, the samples of these races were too small in combination with other factors to provide stable estimates. As these missing demographics are obviously prevalent in Alaska, Hawaii, and Puerto Rico, these areas were excluded from the final analysis. This exclusion is unfortunate as both Native Americans and Puerto Ricans have been shown in past studies to be disproportionately burdened by the asthma epidemic than other groups[11,12].

[10] It should be noted that the statistics presented by the NHIS are based on a sample. These statistics will differ from figures that would be derived from a complete census, or case registry of people in the U.S. with these diseases due to random sampling variability. The results are also subject to reporting, nonresponse and processing errors. These types of errors are kept to a minimum by methods built into the survey. Additionally, a major limitation of is that the information collected represents self-reports of medically diagnosed conditions, which may underestimate disease prevalence since not all individuals with these conditions have been properly diagnosed. However, as noted by ALA September 2004 (footnote #5) the NHIS is one of the best available sources to depict the magnitude of asthma on the national level.

[11] Brody, T.M. (2004, May). Using Small Area Analysis to Estimate Asthma Prevalence in Chicago Public Schools. Public Health GIS News and Information (58), pp. 10-13. Retrieved January 14, 2005 from

Even with the all other races brought into the “Other” category, there was still a shortage of population in the (“65 and Over,” “Other,” “At or Below Poverty Level,” “Male”) and (“65 and Over,” “Other,” “At or Below Poverty Level,” “Female”) cells. In order to acquire a frequency of 30 or more persons in these cells, (“65 and Over,” “Other,” “Males,” “Above Poverty Level,”) and (“65 and Over,” “Other,” “Males,” “At or Below Poverty Level”) were combined as were (“65 and Over,” “Other,” “Females,” “Above Poverty Level,”) and (“65 and Over,” “Other,” “Females,” “At or Below Poverty Level”). Combining these cells resulted in 46 alpha cells for the analysis. See Table 1 for the exhaustive list of these cells.

Identical alpha cells were created from the Census 2000 Summary File 3 (SF 3) for each Census Tract. SF3 presents detailed population and housing data collected from a 1-in-6 sample and weighted to represent the total population. Specifically, the cells were created from SF 3’s PCT75 B through I data. These data present poverty status in 1999 by sex by age for racial groups categorized B through I. B (Black Alone), H (Hispanic), and I (White Alone, Not-Hispanic) were used independently for their corresponding categories while categories, C through G were combined for the “Other” Category.

The population rates of each combination of age, race, gender, and poverty status cell in a given Census Tract were then multiplied by its respective asthma prevalence rate and summed to obtain an estimate for each aforementioned prevalence rate in each Census Tract.

## Results

Table 1 shows the rates derived from the 2000 NHIS for each of the 46 alpha cells. African American male and female children at or below the poverty level appear to have some of the highest rates of all three types of prevalence. Both the African American and Poverty variable appear to drive the prevalence higher throughout the set, although additional statistical tests would be necessary to study these relationships.

The rates by Census Tract are shown graphically in Maps 1-3. In general, one can expect higher prevalence rates in the South and Appalachian Census Tracts as well as major cities. These areas have a greater percentage of African American and impoverished populations associated with higher levels of asthma prevalence.

## Next Steps

It would be interesting to extend this work in several areas. First and foremost, these estimates should be incorporated in decision making tools like those being developed by OECA. Although the estimates should not be considered as accurate as what might be expected from true

http://www.cdc.gov/nchs/data/gis/cdcgis58.pdf

[12] American Lung Association. Lung Disease Data in Culturally Diverse Communities: 2005. Retrieved February 18, 2005 from http://www.lungusa2.org/embargo/lddcdc/LDD.pdf

surveillance, they do provide some indication of where we might expect higher rates of asthma to occur.

Another area of interest is to integrate additional years of NHIS data in order to increase the frequency within demographic categories. More data will likely lead to more stable estimates for groups with small populations in the NHIS. As mentioned before these groups include Puerto Ricans, Native Americans and Alaskan Natives, and Hawaiian Islanders. Including these populations in the analysis would further distinguish sensitivities and allow all 50 States and Puerto Rico to be included in the analysis instead of just the contiguous 48 States.

A third extension of the analysis is to compare and improve the estimated rates with the Behavioral Risk Factor Surveillance System (BRFSS). The American Lung Association believes the BRFSS provides better local estimates of asthma for the adult and senior population than the NHIS[13]. A method incorporating the BRFSS data may provide better localized estimates of the asthma epidemic.

Fourth, it would be ideal if the results could be compared with existing asthma surveillance information. As mentioned before, true surveillance of the asthmatic population does not currently exist. As a surrogate, it may be possible to get the number of emergency room (ER) visits in a given year from State Health Departments. However, emergency room information is only a subset of the most severe prevalence indicator of the three NHIS asthma indicators. NHIS participants are asked if an emergency room *or urgent care center* has been visited in the last year. State Health Departments won't likely have urgent care center data. Additionally, ER data are provided as visits from a particular zip code. Zip codes do not match the same boundaries as Census Tracts. This lack of spatial correlation prevents direct comparisons of ER data with the Census Tract estimates. Even more detrimental for comparison purposes, the ER data would only depict the number of visits from the zip code over a given year, not the number of people visiting. For example, if the data show seven ER visits for asthma in the year 2000 from a given zip code, it may be seven separate individuals, or one person visiting six times and another person visiting once. Without knowing the number of people visiting, it is not possible to create the comparable rates necessary for model validation or effective community comparisons. However the additional information may help support the modeling effort with several assumptions and provide a helpful asthma indicator for decision-making processes.

In closing, the numbers reported in this paper are estimates derived from a modeling effort. Only an effective surveillance program will establish the true rates and variances in these small areas. Some recent initiatives are making it possible for such a network to be developed nationally[14]. Hopefully, in time, a much better understanding of the national asthma epidemic will come from such networks.

[13] American Lung Association Estimated Prevalence and Incidence of Lung Disease By Lung Association Territory, September 2004. Retrieved January 14, 2005 from http://www.lungusa.org/atf/cf/{7A8D42C2-FCCA-4604-8ADE-7F5D5E762256}/ESTPREV2004.PDF

[14] McGeehin, M.A. J.R. Qualters, and A.S. Niskar, National Environmental Public Health Tracking Program: Bridging the Information Gap, Environmental Health Perspectives 112, (14) October 2004, Retrieved February 27, 2005 from http://ehp.niehs.nih.gov/members/2004/7144/7144.html

**Table 1:** Descriptions, Counts, Rates, and Ranks of 46 Alpha Cells Created From NHIS Data

| AGE[15] | SEX[16] | RACE[17] | POV[18] | FREQ[19] | EVER[20] | LST12[21] | ER[22] | PEVER[23] | PLST12[24] | PER[25] | REVER[26] | RLST12[27] | RER[28] |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| C | M | H | Y | 394 | 48 | 20 | 12 | 12.18% | 5.08% | 3.05% | 13 | 18 | 10 |
| C | M | H | N | 1491 | 179 | 76 | 33 | 12.01% | 5.10% | 2.21% | 14 | 17 | 13 |
| A | M | H | Y | 377 | 18 | 4 | 1 | 4.77% | 1.06% | 0.27% | 44 | 43 | 43 |
| A | M | H | N | 1732 | 107 | 22 | 7 | 6.18% | 1.27% | 0.40% | 40 | 42 | 40 |
| S | M | H | Y | 59 | 6 | 2 | 1 | 10.17% | 3.39% | 1.69% | 21 | 31 | 20 |
| S | M | H | N | 174 | 12 | 5 | 3 | 6.90% | 2.87% | 1.72% | 37 | 33 | 19 |
| C | F | H | Y | 409 | 38 | 15 | 5 | 9.29% | 3.67% | 1.22% | 25 | 28 | 27 |
| C | F | H | N | 1328 | 111 | 49 | 20 | 8.36% | 3.69% | 1.51% | 28 | 27 | 23 |
| A | F | H | Y | 626 | 75 | 39 | 21 | 11.98% | 6.23% | 3.35% | 15 | 12 | 7 |
| A | F | H | N | 2069 | 169 | 78 | 31 | 8.17% | 3.77% | 1.50% | 33 | 26 | 24 |
| S | F | H | Y | 105 | 11 | 6 | 5 | 10.48% | 5.71% | 4.76% | 20 | 14 | 4 |
| S | F | H | N | 235 | 23 | 10 | 2 | 9.79% | 4.26% | 0.85% | 23 | 22 | 33 |
| C | M | W | Y | 245 | 45 | 18 | 4 | 18.37% | 7.35% | 1.63% | 2 | 6 | 21 |
| C | M | W | N | 3361 | 465 | 205 | 54 | 13.84% | 6.10% | 1.61% | 11 | 13 | 22 |
| A | M | W | Y | 474 | 72 | 24 | 6 | 15.19% | 5.06% | 1.27% | 7 | 19 | 25 |
| A | M | W | N | 7162 | 576 | 158 | 26 | 8.04% | 2.21% | 0.36% | 34 | 38 | 42 |
| S | M | W | Y | 86 | 5 | 3 | 2 | 5.81% | 3.49% | 2.33% | 42 | 29 | 12 |
| S | M | W | N | 1708 | 144 | 43 | 12 | 8.43% | 2.52% | 0.70% | 27 | 35 | 37 |
| C | F | W | Y | 206 | 22 | 13 | 4 | 10.68% | 6.31% | 1.94% | 19 | 10 | 17 |
| C | F | W | N | 3181 | 310 | 149 | 30 | 9.75% | 4.68% | 0.94% | 24 | 20 | 32 |
| A | F | W | Y | 711 | 109 | 63 | 31 | 15.33% | 8.86% | 4.36% | 6 | 3 | 5 |
| A | F | W | N | 8164 | 919 | 426 | 93 | 11.26% | 5.22% | 1.14% | 18 | 16 | 29 |
| S | F | W | Y | 266 | 22 | 6 | 2 | 8.27% | 2.26% | 0.75% | 31 | 37 | 35 |
| S | F | W | N | 2747 | 227 | 74 | 17 | 8.26% | 2.69% | 0.62% | 32 | 34 | 39 |
| C | M | B | Y | 224 | 43 | 22 | 11 | 19.20% | 9.82% | 4.91% | 1 | 2 | 3 |
| C | M | B | N | 869 | 147 | 60 | 27 | 16.92% | 6.90% | 3.11% | 3 | 9 | 9 |
| A | M | B | Y | 178 | 23 | 7 | 2 | 12.92% | 3.93% | 1.12% | 12 | 24 | 31 |
| A | M | B | N | 1268 | 93 | 30 | 9 | 7.33% | 2.37% | 0.71% | 36 | 36 | 36 |
| S | M | B | Y | 48 | 4 | 1 | 1 | 8.33% | 2.08% | 2.08% | 29 | 40 | 14 |
| S | M | B | N | 198 | 17 | 8 | 4 | 8.59% | 4.04% | 2.02% | 26 | 23 | 15 |
| C | F | B | Y | 230 | 34 | 25 | 18 | 14.78% | 10.87% | 7.83% | 9 | 1 | 1 |
| C | F | B | N | 848 | 121 | 59 | 31 | 14.27% | 6.96% | 3.66% | 10 | 8 | 6 |
| A | F | B | Y | 508 | 83 | 42 | 25 | 16.34% | 8.27% | 4.92% | 4 | 4 | 2 |
| A | F | B | N | 1908 | 191 | 72 | 24 | 10.01% | 3.77% | 1.26% | 22 | 25 | 26 |
| S | F | B | Y | 113 | 17 | 8 | 3 | 15.04% | 7.08% | 2.65% | 8 | 7 | 11 |
| S | F | B | N | 337 | 28 | 11 | 4 | 8.31% | 3.26% | 1.19% | 30 | 32 | 28 |
| C | M | O | Y | 32 | 5 | 2 | 1 | 15.63% | 6.25% | 3.13% | 5 | 11 | 8 |
| C | M | O | N | 266 | 31 | 12 | 2 | 11.65% | 4.51% | 0.75% | 17 | 21 | 34 |
| A | M | O | Y | 58 | 3 | 2 | 1 | 5.17% | 3.45% | 1.72% | 43 | 30 | 18 |
| A | M | O | N | 413 | 26 | 2 | 0 | 6.30% | 0.48% | 0.00% | 39 | 44 | 44 |
| S | M | O | * | 51 | 6 | 4 | 1 | 11.76% | 7.84% | 1.96% | 16 | 5 | 16 |
| C | F | O | Y | 38 | 1 | 0 | 0 | 2.63% | 0.00% | 0.00% | 45 | 45 | 45 |
| C | F | O | N | 254 | 15 | 5 | 1 | 5.91% | 1.97% | 0.39% | 41 | 41 | 41 |
| A | F | O | Y | 89 | 7 | 5 | 1 | 7.87% | 5.62% | 1.12% | 35 | 15 | 30 |
| A | F | O | N | 457 | 31 | 10 | 3 | 6.78% | 2.19% | 0.66% | 38 | 39 | 38 |
| S | F | O | * | 53 | 1 | 0 | 0 | 1.89% | 0.00% | 0.00% | 46 | 46 | 46 |

[15] **AGE:** C = Child (Under 18); A = Adult (18 to 64); S = Senior (65 and Over).

[16] **SEX:** M = Male; F =Female

[17] **RACE:** H = Hispanic; W = White; B = Black; O = Other

[18] **POV:** Y = At or Below Census Poverty Level; N = Above Census Poverty Level

[19] **FREQ** = Frequency

[20] **EVER:** Yes, a doctor or other health professional EVER told you that … had asthma.

[21] **LST12:** Yes, during the past 12 months, … have (has) had an episode of asthma or an asthma attack.

[22] **ER:** Yes, during the past 12 months, … did have to visit an emergency room or urgent care center because of asthma.

[23] **PEVER:** Percent answering yes, a doctor or other health professional EVER told you that… had asthma.

[24] **PLST12:** Percent answering yes, during the past 12 months, … have (has) had an episode of asthma or an asthma attack.

[25] **PER:** Percent answering yes, during the past 12 months, … did have to visit an emergency room or urgent care center because of asthma.

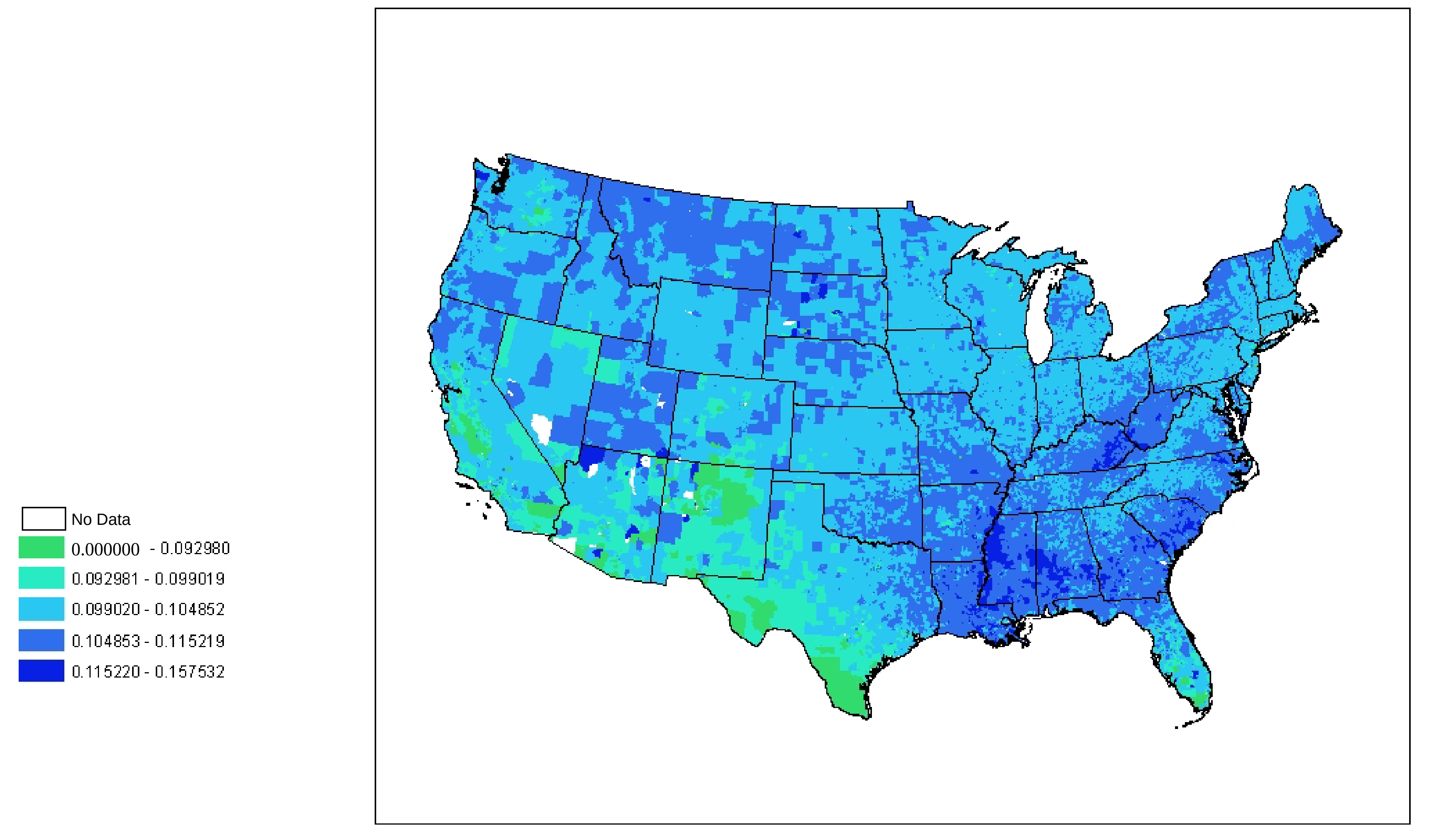
[26] **REVER:** Rank of percent answering yes, a doctor or other health professional EVER told you that… had asthma (1= highest).

[27] **RLST12:** Rank of percent answering yes, during the past 12 months, … have (has) had an episode of asthma or an asthma attack (1= highest).

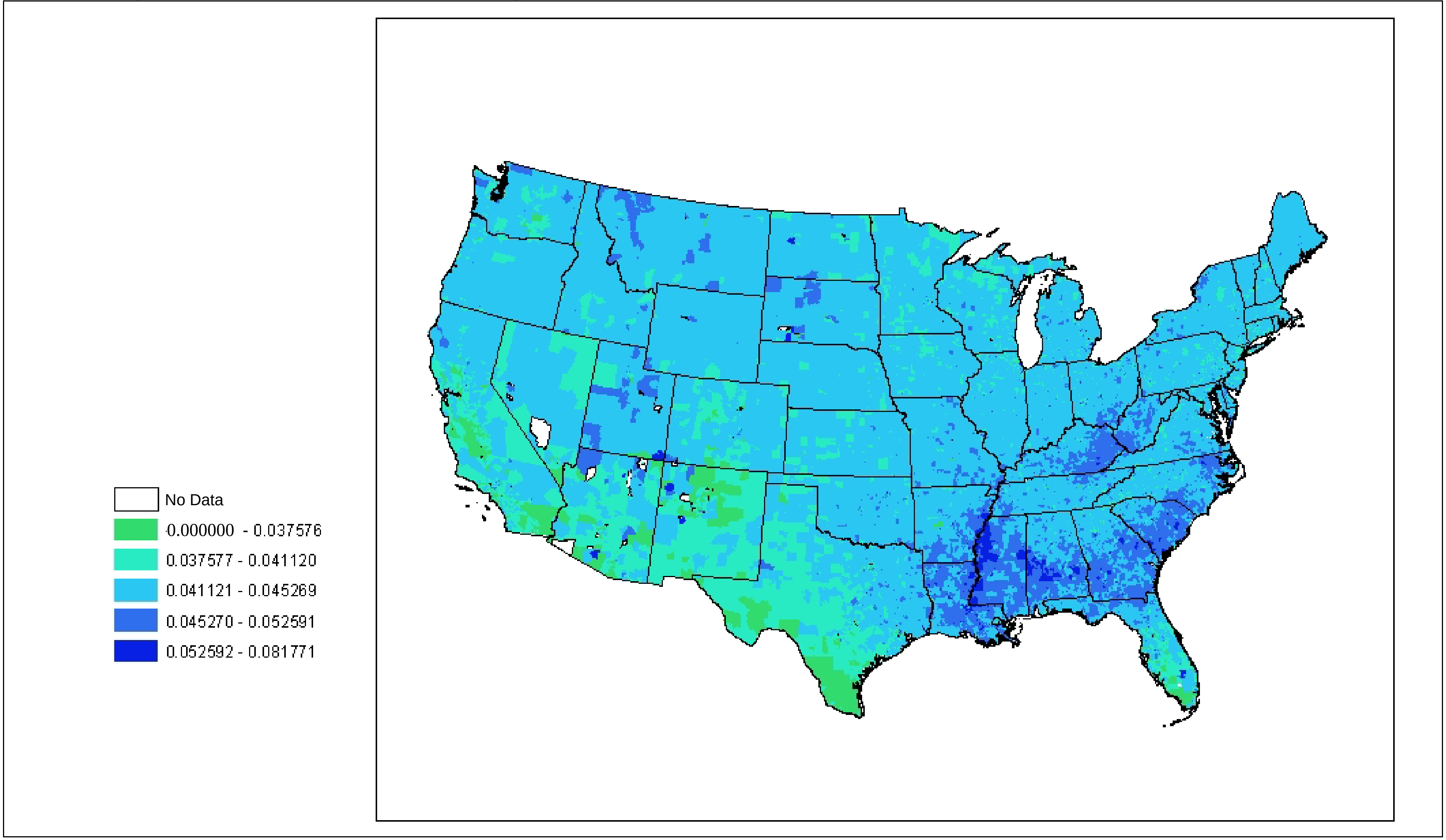
[28] **RER:** Rank of percent answering yes, during the past 12 months, … did have to visit an emergency room or urgent care center because of asthma (1= highest).

* Senior, Other, Males "Above" and "At or Below Poverty" were combined as were Senior, Other, Females "Above" and "At or Below Poverty" due to low frequency in the survey.

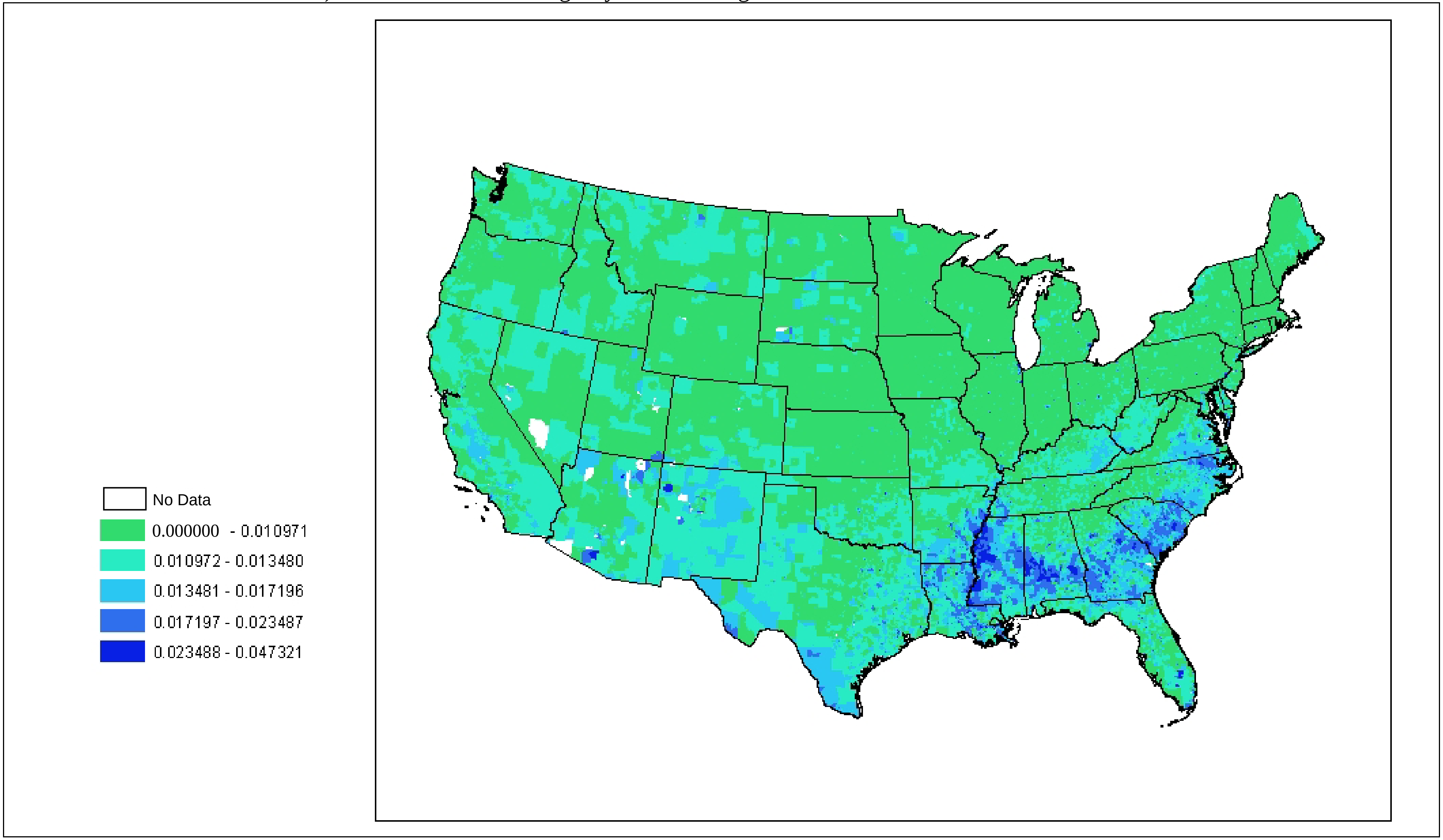
**Map 1:** Estimated Percentage of Census Tract Population Answering “Yes” To The Question “Has A Doctor or Other Health Professional EVER Told You That You (or The Sample Child In The Household) Had Asthma?”



**Map 2:** Estimated Percentage of Census Tract Population Answering “Yes” To The Question “During The Past 12 Months, Have You (or The Sample Child In The Household) Had An Episode of Asthma or An Asthma Attack?”



**Map 3:** Estimated Percentage of Census Tract Population Answering "Yes" To The Question "During the past 12 Months, Did You (or The Sample Child In The Household) Have to Visit An Emergency Room or Urgent Care Center Because of Asthma?"



## The SAS System for Analyzing Binary Responses

**John Bander, SAS, Public Sector Group**

Logistic regression enables you to investigate the relationship between a categorical outcome and a set of explanatory variables. It has become a workhorse of modern analytical work. This talk will survey the wealth of options available in the SAS System for performing logistic regression. Options and syntax will be illustrated with code samples.

# Bayesian Estimation of the Normal Mean in Presence of Non-Detects

by
Ahmed Khago
Department of Mathematical Sciences
University of Nevada, Las Vegas

## Table of Contents

## Summary

The presence of non-defects data in Environmental applications is very common practice and makes it difficult to decide which method is appropriate to incorporate these data in estimating the population parameters and they impact the upper confidence limit of the mean which is required for many remediation decisions.

This paper is concern with the Bayesian approach to estimate the mean when encountered with left-censored data sets. Considering the joint non-informative prior, we derived the posterior probability density function of the mean of left-censored data. However, this density function is not recognizable and we do not know analytically the constant that would normalize the density function such that the integral would equal to 1. In other words, we do not know analytically the posterior moments. Numerical integration using adaptive Simpson quadrature rule function in matlab to obtain the numerical posterior mean and the upper confidence limit (UCL). Several numerical examples are given which illustrate the practical application of these results.

## Section 1

## Introduction

In environmental sciences application, many data measurements such as herbicide concentration in soil, air, and water do not get reported because such measurements fall below a certain detection limit ("DL") and many groundwater monitoring applications of the United States Environmental Protection Agency ("EPA") do not require reporting such data. These measurements, however, cannot be ignored since they impact the upper confidence limit ("UCL") of the mean which is required for many remediation decisions. The DL is the lowest level of concentration of any particular substance that can be reliably detected and is statistically different from a "blank" reading.

In most environmental applications, this non-reported data, with observations recorded as being below a certain limit, is called "censored" data – which means the observations are not available at one or both ends. Censoring usually occurs when the pollutant concentration is very near or below the DL; however, this practice creates special problems and makes it difficult to analyze and summarize data sets and could lead to biased estimations of the population parameters, such as the mean and the standard deviation.

Censored data are classified into four major ways: truncated vs. censored, left vs. right, single vs. multiple, and censored type I vs. censored type II (Cohen 1991, pp.3-5). Environmental Science applications mostly deal with type I left censored.

A data sample is said to be left truncated if the truncation point ("T") is known and the value of the observations below T is deleted or not reported, but the values above T are known and are reported. For example, consider the data set: 3, 4, 3, 5, 4, 3, <2, 2, 3, <2, <2, with a DL of 2. All the data values reported as "<2" will be eliminated and if no indication of how many observation were excluded, this would be called a type I, left—truncated sample. On the other hand, a data set of size "n" is said to be left—censored if the censoring level T is known and the value of the observations below T level is known (k observations) only to fall below T while the known observations above T level are fully known and reported (n-k observations). For the above example, the values reported as "<2" will not be eliminated.

The difference between truncated data and censored data is that the censored data points are those whose measured properties are not known precisely, but are known to fall above or below some DL, or limiting sensitivity. On the other hand, truncated data points are those which are missing from the sample altogether due to sensitivity limits.

The most common method of dealing with censored data in environmental sciences is the substitution method. One way is to replace the non-detect values with zero or deleting the censored data. The reason behind the use of this method is that interest is always in the detected data. This method produces biased results because the statistics test only applies to the detected data. A second way is to replace each censored observation with an arbitrary fraction of the DL. The most common substitution is to replace the censored data by half the detection limit or by the detection limit itself. Singh and Nocerino (2002) pointed out the replacement of half the detection limit produced biased estimate of mean and error increases when multiple detection limits are present.

Another approach is the maximum likelihood estimation ("MLE"), which is often used in environmental studies. There are three types of information needed to perform the calculation: the values of data above detection limits, the proportion of data below detection limits, and the parametric form of the assumed distribution. For small data sets, however, the MLE would perform poorly (Gleit, 1985; Shumway, et al, 2002). MLE is an efficient method to estimate the parameters when the data set is large enough. MLE is performed by solving the maximum likelihood function ("L") for the parameters $\mu$ and $\sigma$ and by taking the natural log and maximizing the Ln (L) by setting the derivative equal to zero and solving for the parameters.

The third approach involves non-parametric procedures, which are called the distribution-free methods and are commonly used in environmental sciences. These methods are useful for censored data because they use the available information; however, data with multiple thresholds are still not familiar to most environmental scientists.

Researchers from various disciplines studied the estimation of the parameters of the normal populations from censored samples. One of these researchers is Cohen [1950-1959]. He derived the maximum likelihood estimation ("MLE") for determining the mean and the standard deviation of censored data. Cohen's MLE uses both the detected observation and the proportion of data set below detection limits to compute and analyze statistics for the entire data set. The MLE method requires that the distribution of the data to be known and specified. In environmental sciences, the normal and lognormal distributions are usually used. The MLE equation is then solved using numerical methods, such as the Newton-Raphson method; however, the MLE method has been shown to perform poorly with data set less than 25 to 50 observations (Gleit, 1985; Shumway, et al, 2002).

Gilbert and Kinnison (1981) studied and evaluated the methods of substitution, deleting censored data and Cohen's table lookup. They concluded that substituting for a detection limit is biased. Gleit (1985) found MLE did not perform well for a small data set, even though the assumed distribution is known. He concluded MLE methods work poorly for small sample sizes and the substitution method of detection limits also worked poorly. Gillion and Hesel (1986) found that the MLE method worked well when the assumed distribution matched that of data. They also found that the substitution method worked poorly. Gilbert (1987) considered several methods to calculate an unbiased estimate of the sample mean. The data set should be sampled form normal or lognormal distributions and should include censored data. The data set then should be sorted out and ordered and with an equal number of observations can be deleted. The trimmed mean can be calculated from these values. The trimmed mean is usually recommended to estimate the mean of a symmetric distribution, even if the data set does not have missing values.

Another method is called "winsorizing" the data set and is considered by Dixon and Tukey (1968), in which we replace the data set in both ends of the data series with the next extreme value in both ends and compute the mean of the new data. The difference between the trimmed mean method and the winsorized method is the trimmed method discards data on both ends of the data set and computes the mean of the remaining data; but the winsorized method replaces data in both ends with the next most extreme datum in each end and then computes the mean of the new data set. Winsorization can be used to estimate the mean and the standard deviation of a symmetric distribution, even though the data set has missing values at one or both ends of the ordered data set.

Our goal in this paper is to compute the Bayes estimate of the mean of a normal population when the data set has non-detects. We present several examples using simulated data and compute the Bayes estimate obtained from left-censored samples with that obtained from the uncensored samples.

## Section 2

In this section, we will discuss some of the popular methods to estimate the mean and variance of a population when only censored data is available. Some of these methods are the trimmed mean, the winsorized mean, and the maximum likelihood.

In situations where non-detect values are reported even when the measurements are below the detection limit, the population mean $\mu$ and the variance $\sigma^2$ can be estimated by calculating the sample mean $\overline{x}$ and the sample variance $s^2$ using one of the following :

1. Calculate $\overline{x}$ and $s^2$ using the full data set, including non-detect values.
2. Delete all non-detects and calculate only $\overline{x}$ and $s^2$ using only the detected data set.
3. Replace every non-detect values with zero and then calculate $\overline{x}$ and $s^2$.
4. Replace the non detected values with values generated from uniform over $[0, DL]$ then calculate $\overline{x}$ and $s^2$.

All of the methods mentioned above are biased estimators, except the last one if the measurements between zero and detection limit are uniformly distributed.

## The trimmed mean

The trimmed mean is one of the methods of estimating the mean of symmetric distribution and it's a compromise between the median and the mean. Several of the lowest and the highest observations are trimmed off (np observations), where $0 < p < .5$, and then the mean of what is left off $(n(1-2p))$ is calculated. Common trimming is 25% of the data at each end. The resulting mean of the central 50% of data is commonly called the "trimmed mean." For example, suppose n=25, data collected from a symmetric distribution has a true mean $\mu$. We can estimate $\mu$ using a 25% trimmed mean. We first compute $.25n = .25(25) = 6.25$. Hence, we can discard the 6 smallest and the 6 largest data. The mean of the remaining is 25-12=13; data is the estimate of the mean.

## The winsorized mean

The use of the winsorized mean method is also one of the recommended methods to estimate the mean of censored data of symmetric distribution. Details of this method are given by Dixon and Tukey (1968).

Given N data set and k non-detect values, the winsorized procedures are as follows:

1. Replace the k non-detects values by the next datum.
2. Replace the k largest values by the next smallest datum.
3. Calculate the sample mean $\overline{x}_w$ and standard deviation S of resulting N data.
4. The resulting estimate $\overline{x}_w$ is unbiased estimator of $\mu$.

The following sample from a well represents the concentration for hazardous chemicals ordered from the smallest to largest. Trace, trace, trace, .67, 2.4, 3.1, 3.5, 3.9, 4.1, 4.6, 5.7, 6.9, 7.5, and 9.1. Replace the three trace concentrations by .67 and the three largest concentrations by 5.7. The data becomes .67, .67, .67, .67, 3.1, 3.5, 3.9, 4.1, 4.6, 5.7, 5.7, 5.7, and 5.7. The sample mean of the new data is $\overline{x}_w$=3.36. This $\overline{x}_w$ is the winsorized mean.

## Bootstrap method

Bootstrap method is a non-parametric method that requires no assumptions regarding the population such as the normal assumption. These techniques are used to reduce the bias in point estimate and build a confidence interval for any parameter. It's a form of a larger class of methods that resample from the original data set and therefore are called resampling procedures. We can obtain accurate confidence intervals without having to make normal theory assumption and estimate the distribution Z directly from the data set. The procedure is described as follows:

Let $x_1, x_2, ........., x_n$ be a random sample of size n, then B bootstrap samples are generated from the original data set. Each bootstrap sample should have n elements, which is generated by sampling with replacement n times. Bootstrap replicates $\overline{X}_1, \overline{X}_2, ..............., \overline{X}_b$ are calculated from the replicates. We can also calculate the estimate $\overline{X}_B$ , the bootstrap standard error , $s_B = \sqrt{\dfrac{\sum(\overline{X}_i - \overline{X}_B)^2}{n-1}}$, and obtain the confidence interval using $z = \dfrac{\widehat{\theta} - \theta}{s_B}$. Finally, $(1-\alpha)100\%$ confidence interval for $\theta$ is $\left(\widehat{\theta} - z_\alpha s_B, \widehat{\theta} + z_\alpha s_B\right)$.

## Section 3

One of the most difficult and controversial problems in environmental data analysis is deciding the appropriate method of incorporating the censored data in computing summary statistics, corresponding tests of hypotheses, and interval estimation of parameters. This is mostly because the choice of method depends on the degree censoring (for example, 10% versus 90% non-detects) and this also depends on the type of application. Most of the methods are available to replace the detection limit with an arbitrary constant. These methods use the probability theory to estimate the shape of the tail of the population density function that was censored and assume the distribution of the sampled population is known; however, if the sample size is small and the censored data percentage is very high, it becomes very difficult to determine the population distribution from the sample. This paper will provide a Bayesian estimate of the mean from left censored data set.

Left-truncated normal distribution has been utilized by a variety of disciplines, such as environmental sciences, economics and finance. Pearson and Lee (1908), Fisher (3), Hald (1949), and Cohen studied singly truncated normal samples when the truncation point is known and the sample size of unmeasured observations is unknown. Stevens (1938), Cochran (1949) and Hald ( 1949) studied singly truncated normal samples when the truncation point is known and the sample size of unmeasured observation is known.

Consider a random variable $X$ from a normal distribution with a probability density function $f(x)$ specified as:

$$f(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, -\infty < x < \infty \qquad (1)$$

The following are graphs of the standardized normal distribution function and cumulative standardized normal curve. The curve is symmetric around $z = 0$.

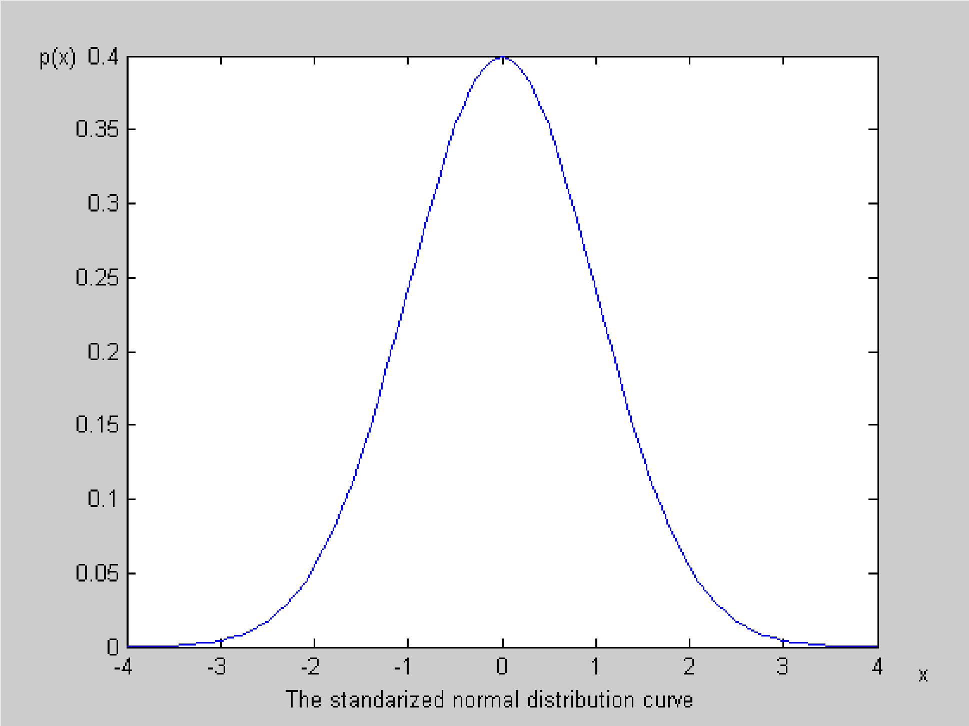


**Figure 1. Standardized Normal Distribution**

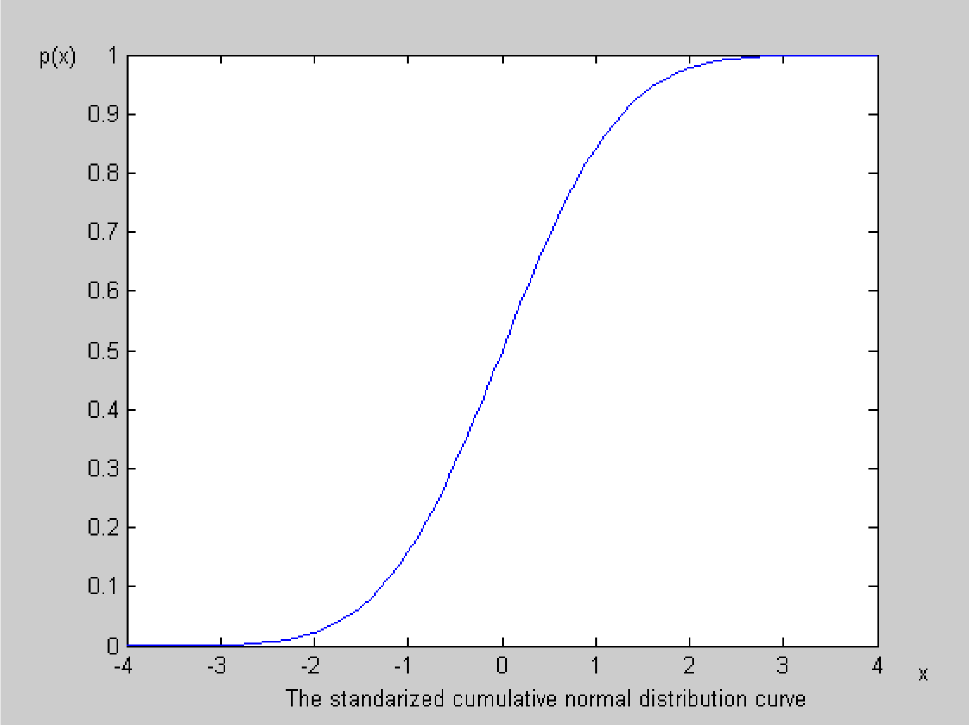


**Figure 2 . Standardized cumulative normal distribution**

## The posterior density of the mean of censored data.

Let $x_1, x_2, x_3, ..., x_n$ be a random sample from a normal distribution $N(\mu, \sigma)$ and suppose k of these measurements falls below the detection limit, DL. Let $\phi$ be the probability density function ("pdf") and $\Phi$ be the cumulative density function ("cdf"), then the likelihood function is the following (Persson and Rootzen 1977):

$$L(x, \mu, \sigma) = [\Phi(z)]^k (2\pi\sigma)^{-(n-k)/2} \exp-[\sum_{i=k+1}^{n} (y_i + z\sigma)^2 / 2\sigma^2] \qquad (1)$$

where $\Phi(z)$ = the probability of an observation less than detection limit, DL and k= the number of observation below detection limit. $\Phi(z)$ can be written as the cumulative density function:

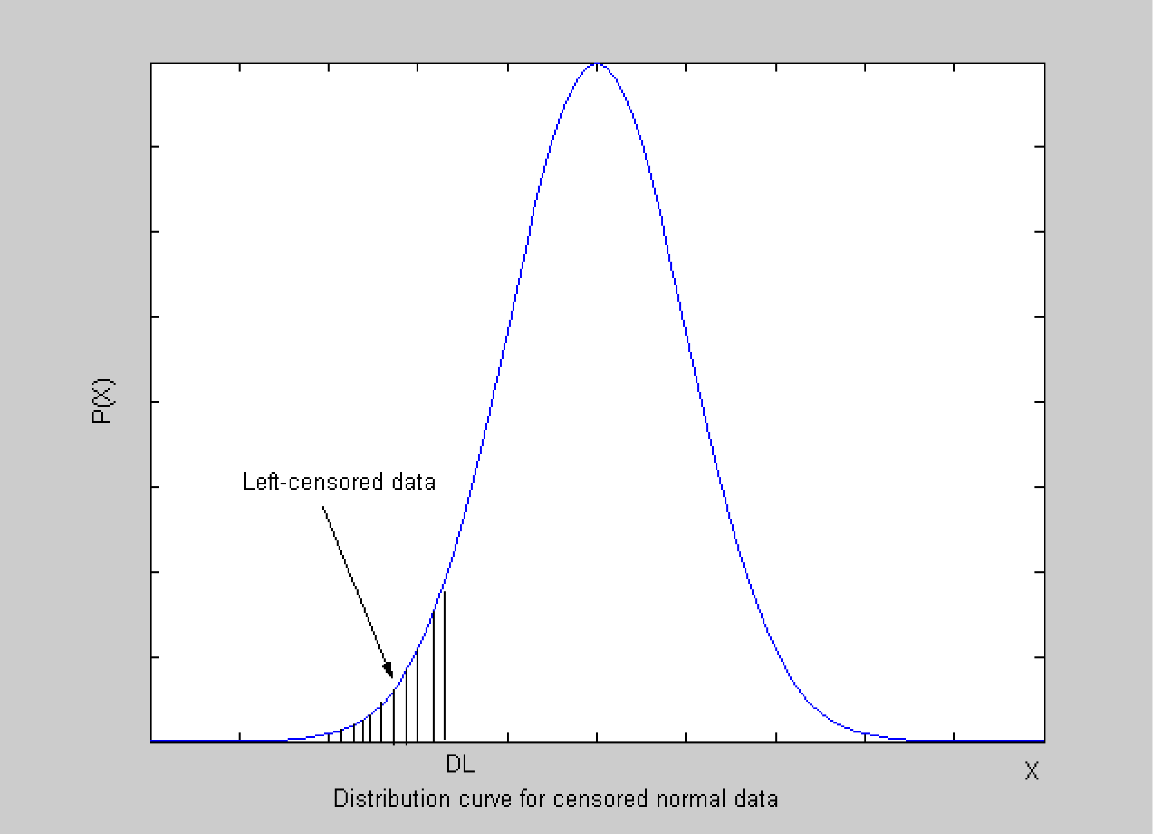


**Figure 3 . Left-censored normal distribution**

$$[\Phi(z)]^k = \left\{ P\left( \frac{X-\mu}{\sigma} < \frac{DL-\mu}{\sigma} \right) \right\}^k \qquad (2)$$

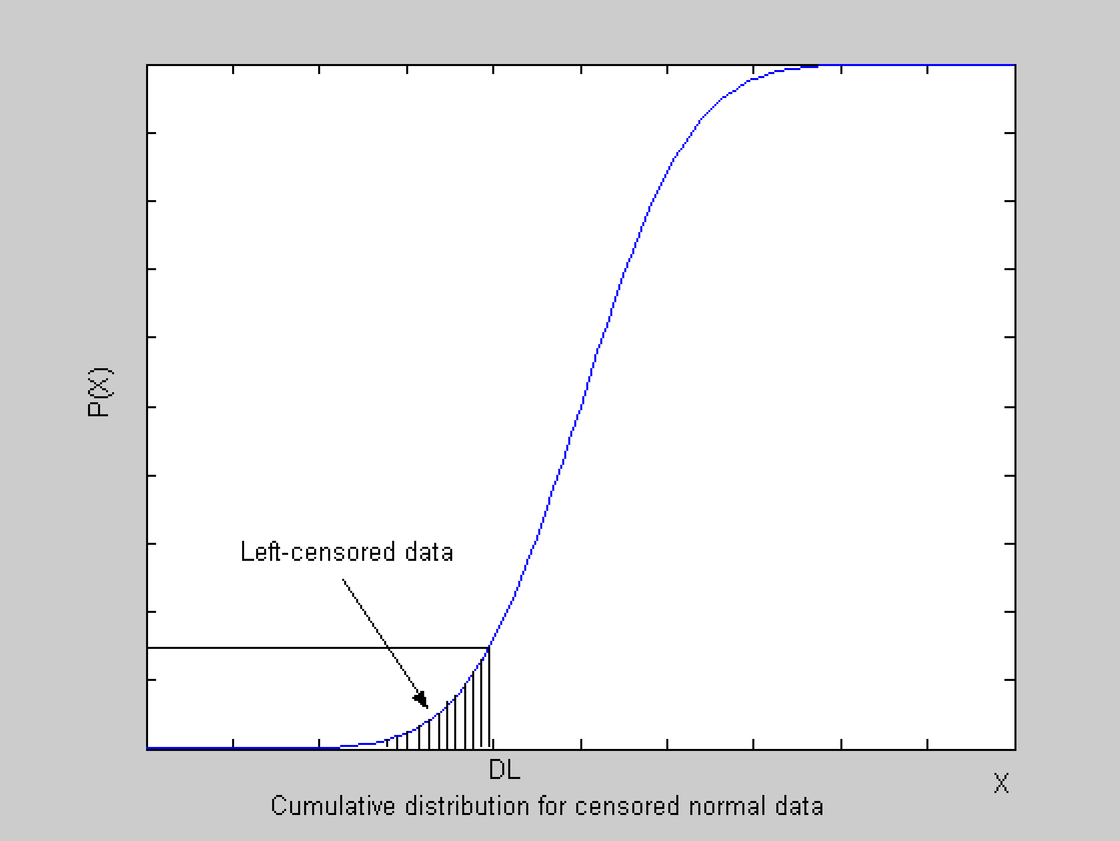


**Figure 4. Cumulative censored normal distribution**

$$\Phi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\frac{DL-\mu}{\sigma}} e^{\left(\frac{x-\mu}{\sigma}\right)^2} dx \qquad (3)$$

$$[\Phi(z)]^k = \left\{ 1 - \frac{1}{\sqrt{2\pi}} \int_{\xi}^{\infty} e^{-\frac{z^2}{2}} dz \right\}^k \qquad (4)$$

Where $\xi = \frac{X-DL}{\sigma}$

The integration of the normal distribution is possible and easier using what we call the error function. The error function is twice the integral of the standardized normal distribution with $\mu = 0$ and $\sigma = 1$. The error function is defined as:

$$erf(x) = \frac{2}{\sqrt{\pi}} \int_0^{\infty} e^{-u^2} du \qquad (5)$$

$$erfc(x) = 1 - erf(x) = \frac{2}{\sqrt{\pi}} \int_x^{\infty} e^{-u^2} du \qquad (6)$$

Using the error function, we can write the $\Phi(z)$ in simpler form;

$$[\Phi(z)]^k = \left\{P\left(\frac{X-\mu}{\sigma} < \frac{DL-\mu}{\sigma}\right)\right\}^k = \left[\int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma}} e^{\left(\frac{DL-\mu}{\sigma}\right)^2} dx\right]^k \qquad (7)$$

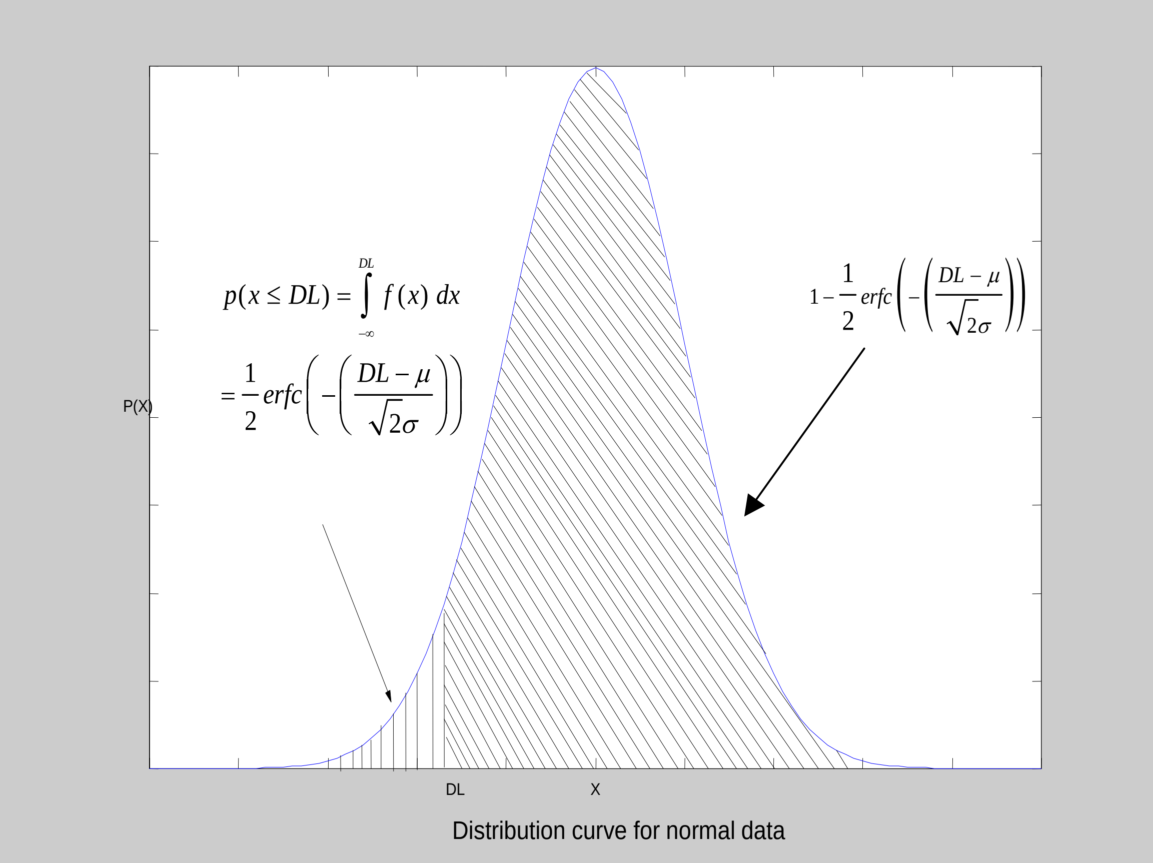


**Figure 5. Normal distribution in term of error function**

Let $u = \left( \dfrac{DL - \mu}{\sqrt{2}\sigma} \right)$,and simplify. The integral becomes:

$$\int_{-\infty}^{\left(\frac{DL-\mu}{\sqrt{2}\sigma}\right)} \frac{1}{\sqrt{\pi}\sigma} e^{-u^2} du = \frac{1}{2} erfc\left( -\left( \frac{DL - \mu}{\sigma} \right) \right) \tag{8}$$

$$[\Phi(z)]^k = \left[ \frac{1}{2} erfc\left( -\left( \frac{DL - \mu}{\sigma} \right) \right) \right]^k \tag{9}$$

As prior for $\mu$ and $\sigma$ , we will use the non-informative prior:

$$g(\mu, \sigma) = 1/\sigma^2 \tag{10}$$

Combining this prior with the likelihood function yields the posterior pdf for $(\mu, \sigma)$ :

$$g^*(\mu, \sigma \mid x) \propto f(x \mid \mu, \sigma) g(\mu, \sigma) \tag{11}$$

$$g^*(\mu, \sigma \mid x) \propto [\Phi(\mathrm{z})]^k (1/\sigma^2)(1/\sigma^2)^{(n-k)/2} \exp - [\sum_{i=k+1}^{n} (x_i - \mu)^2 / 2\sigma^2] \tag{12}$$

It is not possible to analytically integrate out $\mu$ and $\sigma$ from this probability density function to obtain the marginal posterior pdf's: $g^*(\mu \mid x)$ and $\mathrm{g}^*(\sigma^2 \mid x)$ . Also, the conditional posterior pdf's: $g^*(\mu \mid x)$ and $\mathrm{g}^*(\sigma^2 \mid x)$ are not recognizable densities. In other words, we do not know analytically the constant $K(x)$, such that $g^*(\mu \mid x) / K(x)$ is a properly normalized density, i.e. such that $\int g^*(\mu \mid x) dx = 1$.

The marginal truncated distribution for $\mu$

$$g^*(\mu \mid x) \propto [\Phi(z)]^k \exp - [n(\mu - \overline{x})^2 / 2\sigma^2] \tag{13}$$

The posterior truncated mean is given by

$$\mu_{TP} = E(x) = \int_{-\infty}^{\infty} x \, g^*(\mu \mid x) \, dx \tag{14}$$

$$\propto \int_0^\infty x[\Phi(z)]^k (1/\sigma^2)(1/\sigma^2)^{(n-k)/2} \exp-[\sum_{i=k+1}^n (x_i-\mu)^2/2\sigma^2]\, dx$$

This posterior is not recognizable density; therefore numerical integration will be implemented to find the estimate of the posterior mean.

## The posterior density of the mean of uncensored data

Let $x_1, x_2, x_3, ..., x_n$ be a random sample from a normal distribution $N(\mu,\sigma)$, then the likelihood function is the following:

$$L(x,\mu,\sigma) = (2\pi\sigma)^{-n/2} \exp-[\sum_{i=1}^n (y_i + z\sigma)^2/2\sigma^2] \tag{15}$$

Using the same joint prior, the joint posterior density:

$$g^*(\mu,\sigma \mid x) \propto (1/\sigma^2)^{(n+2)/2} \exp-[\sum_{i=1}^n (x_i-\mu)^2/2\sigma^2] \tag{16}$$

$$= (1/\sigma^2)^{(n+2)/2} \exp-\frac{1}{2\sigma^2}[\sum_{i=1}^n (x_i-\overline{x})^2 + n(\mu-\overline{x})^2] \tag{17}$$

Where $\overline{x}$ is the sample mean of $x_i$. The conditional pdf's from the equation above are:

$$g^*(\mu \mid \sigma^2, x) \propto \exp\left\{\frac{n}{2\sigma^2}\left(\mu-\overline{x}\right)^2\right\} \tag{18}$$

$$g^*(\sigma^2 \mid \mu, x) \propto (1/\sigma^2)^{(n+2)/2} \exp-[\sum_{i=1}^n (x_i-\mu)^2/2\sigma^2] \tag{19}$$

$$g^*(\mu \mid \sigma^2, x) = N(\overline{x}, \sigma^2/n) \tag{20}$$

The posterior mean is given by:

$$E(\mu_P) = \int_{-\infty}^\infty g^*(\mu \mid x) d\mu_P \tag{21}$$

$$\propto \int_{0}^{\infty} x(1/\sigma^2)(1/\sigma^2)^{n/2} \exp-[\sum_{i=1}^{n}(x_i-\mu)^2/2\sigma^2]\, dx \qquad (22)$$

This is the probability posterior density function for the mean of uncensored data. Numerical examples will be illustrated in the next chapter.

## Section 4

## Examples

**Example 1A (Complete Data Set)**

A simulated data set of size 30 was generated from a normal population with mean, $\mu$ =1 and $\sigma$ =1, N (1, 1). 0.33483, 1.07417, 0.91798, 0.53191, 1.58731, 2.72819, 0.95847, 1.7179, 0.18525,0.43238, 1.35569,1.95343, 0.93426, -0.46753, -0.2097, 0.33177, 2.63655,-0.10443, 2.48921,3.87581, 1.98537, 0.01594, 1.01421, .13981, 2.16441, 1.6618, 3.80945, 0.40988, 1.41659, 1.22999.

The sample mean and the standard deviation using the full uncensored data were 1.27 and 1.099, respectively. The following is the plot of the posterior density of the mean, $g^*(\mu|\sigma^2,x)$, using Mathematica. The posterior mean estimate, using equation (18) the mean, is 1.307.

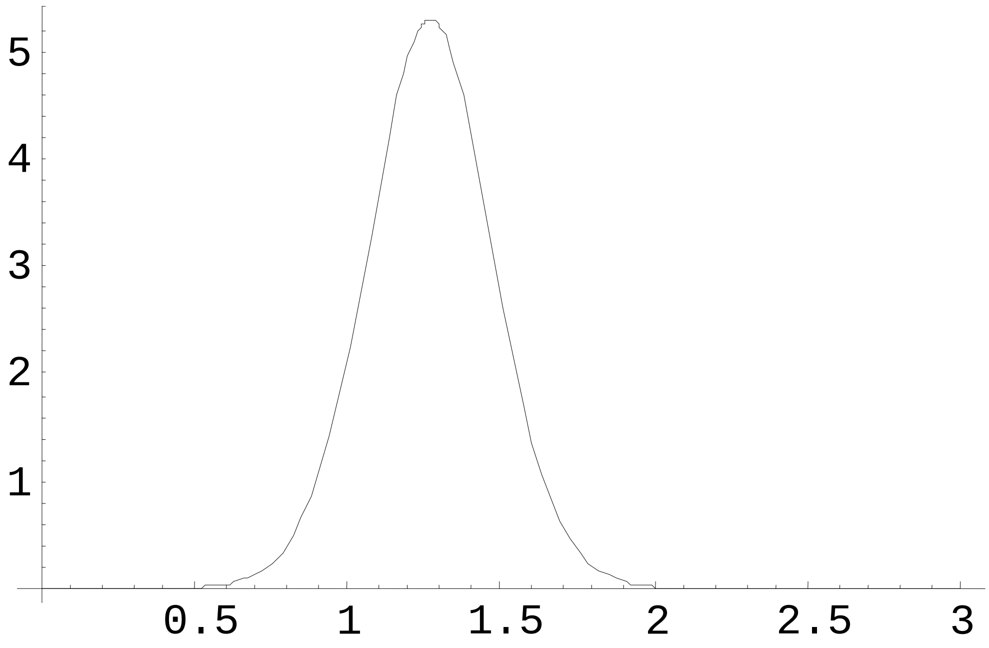


**Figure 5, Posterior Density of the mean,** $\mu = 1.27$, $\sigma = 1.099$

**Example 1B (Censored Data)**

A simulated data set of size 30 was generated from normal population with mean, $\mu$ =1 and $\sigma$ =1, N (1, 1) with DL=.1 and k=4. The left censored data are: <.1,<.1, <.1, <.1, 0.33483, 1.07417, 0.91798, 0.53191, 1.58731, 2.72819, 0.95847, 1.7179, .18525, 0.43238, 1.35569, 1.95343, 0.93426, 0.33177, 2.63655, 2.48921, 3.87581, 1.98537, 1.01421, 1.13981, 2.16441, 1.6618, 3.80945, 0.40988, 1.41659, and 1.22999. The sample mean and the standard deviation obtained using the 26 observed values were: 1.49 and 1.001, respectively. The following is the plot of the posterior density of the mean: $g^*(\mu \mid \sigma^2, x)$, using Mathematica. The posterior mean estimate using equation (14) is 1.5091 and the upper confidence level (UCL) is 1.875.

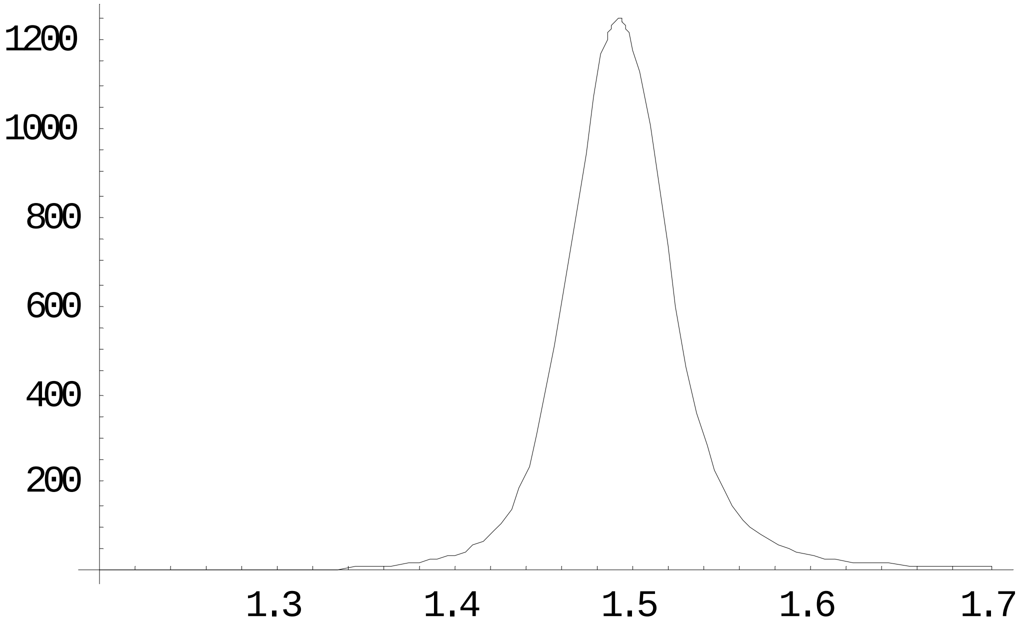


**Figure 6, Posterior Density of the mean,** $\mu = 1.49$, $\sigma = 1.001$

**Example 2A( Full Data)**

A simulated data set of size 30 was generated from a normal population with mean, $\mu$ =1 and $\sigma$ =1, N (1, 1).-0.34637, 1.23544, 0.65759, 0.55156, 0.73505, 2.30196, 0.44569, 1.87822, 1.274, 0.95734, 1.10993, 0.10149, 0.89895, 2.13774, 1.35832, 1.30284, 1.99124, 0.20874, -0.64009, -1.57503, 1.51805, 2.0091, 2.60781, -0.56341, 1.34461, 0.88987, 0.20914, -0.4529, 1.76152, and 0.18125.

The sample mean and the standard deviation using the full data were: 0.870 and 0.988, respectively. The following is the plot of the posterior density of the mean: $g^*(\mu | \sigma^2, x)$, using Mathematica. The posterior mean estimate using equation (18) is 0.967.

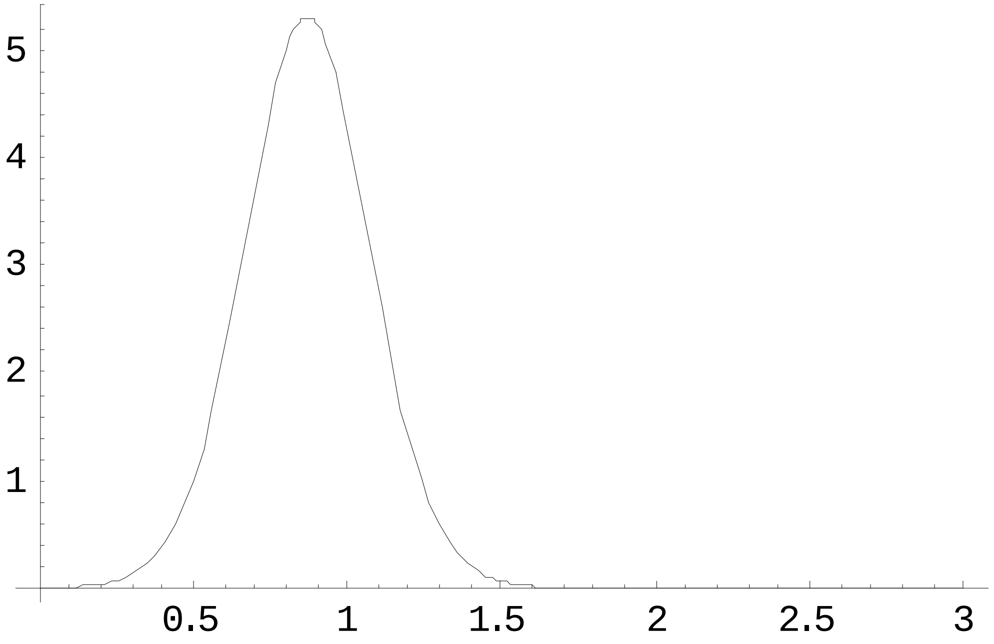


**Figure 7. Posterior Density of the mean,** $\mu = .87,\ \sigma = .988$

## Example 2B (Censored Data)

A simulated data set of size 30 was generated from a normal population with mean, $\mu$ =1 and $\sigma$ =1, N (1, 1) with L=.1 and k=5. The left censored data are: <.1, <.1, <.1, <.1, <.1, 1.23544, 0.65759, 0.55156, 0.73505, 2.30196, 0.44569, 1.87822, 1.274, 0.95734, 1.10993, 0.10149, 0.89895, 2.13774, 1.35832, 1.30284, 1.99124, 0.20874, 1.51805, 2.0091, 2.60781, 1.34461, 0.88987, 0.20914, 1.76152, and 0.18125. The sample mean and the standard deviation obtained using the 25 observed values were 1.187 and 0.715, respectively. The following is the plot of the posterior density of the mean: $g^*(\mu|\sigma^2, x)$, using Mathematica. The posterior mean estimate using equation (14) is 1.2565 and the upper confidence level (UCL) is 1.33.

.

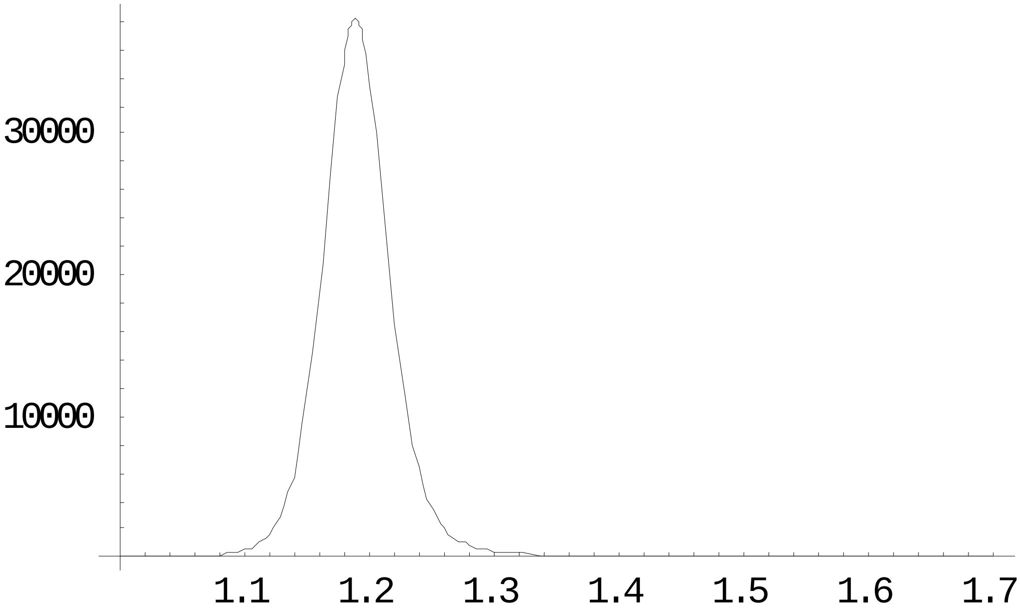


**Figure 8. Posterior Density of the mean,** $\mu = 1.187, \ \sigma = .715$

**Example 3A ( Full Data)**

A simulated data set of size 30 was generated from normal population with mean, $\mu$ =1 and $\sigma$ =1, N (1, 1).1.1509, 2.33669, 3.03262, 1.41328, 2.12521, 2.18608, 1.91767, 1.42204, 1.78774, -0.98267, 2.60349, -0.53495, 0.53363, 1.59111, 0.8585, 0.17489, 2.23636, 0.18885, 1.38405, 1.23115, 0.54023, 2.40644, 0.3547, 1.25482, 0.98104, 0.63982, 1.56435, 1.33922, 1.03252, and 1.33326.

The sample mean and the standard deviation using the full data were 1.27 and 0.921, respectively. The following is the plot of the posterior density of the mean: $g^*(\mu \mid \sigma^2, x)$, using Mathematica. The posterior mean estimate using equation (18) is 1.307.

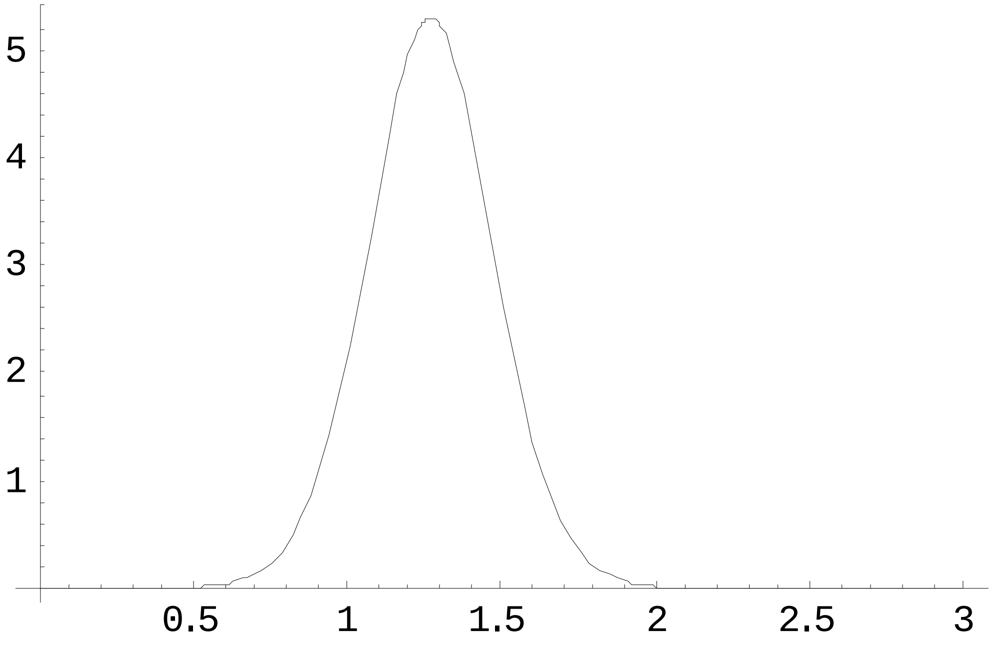


**Figure 9. Posterior Density of the mean,** $\mu = 1.27$, $\sigma = .921$

## Example 3B(Censored Data)

A simulated data set of size 30 was generated from normal population with mean, $\mu$ =1 and $\sigma$ =1, N (1, 1) with L=.1 and k=2. The left censored data are: <.1, <.1, 1.1509, 2.33669, 3.03262, 1.41328, 2.12521, 2.18608, 1.91767, 1.42204, 1.78774, 2.60349, 0.53363, 1.59111, 0.8585, 0.17489, 2.23636, 0.18885, 1.38405, 1.23115, 0.54023, 2.40644, 0.3547, 1.25482, 0.98104, 0.63982, 1.56435, 1.33922, 1.03252, 1.33326. The sample mean and the standard deviation obtained using the 28 observed values were 1.415 and 0.750, respectively. The following is the plot of the posterior density of the mean: $g^*(\mu \mid \sigma^2, x)$, using Mathematica. The posterior mean estimate using equation (14) is 1.5256 and the upper confidence level (UCL) is 1.18.

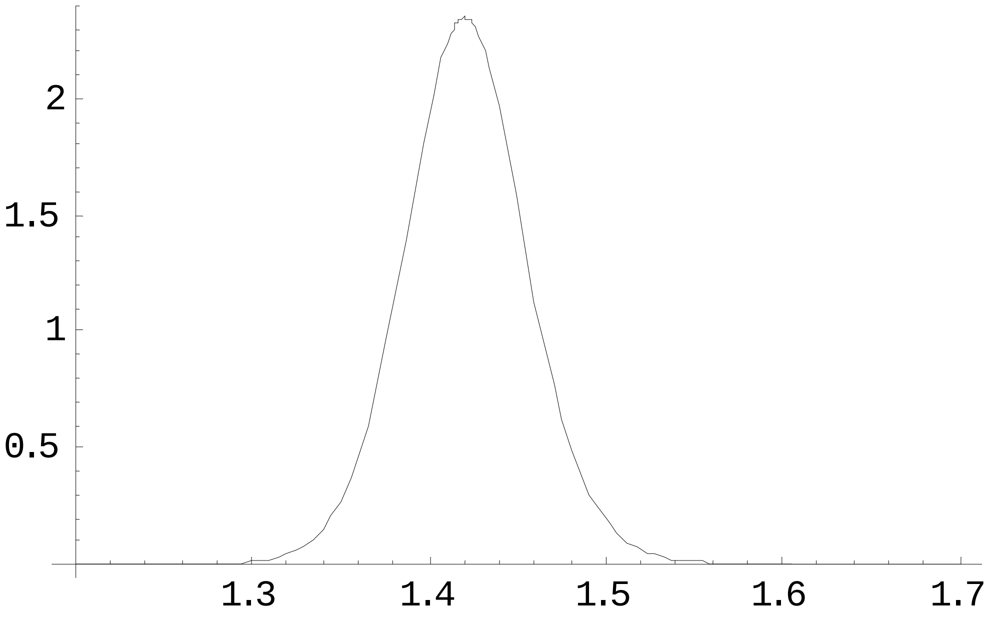


**Figure 10. Posterior Density of the mean,** $\mu = 1.415,\ \sigma = .750$

**Example 4A ( Full Data)**

A simulated data set of size 30 was generated from normal population with mean, $\mu$ =1 and $\sigma$ =1, N (1, 1), 0.11126, 1.75206, 1.17052, 3.15024, 3.18094, 1.56179, 0.927, 2.14169, -0.46995, 2.18118, 0.98145, 1.41042, 3.10198, 2.78779, 0.71599, 0.54362, 0.5441, 2.71058, 2.60982, 0.77772, 1.80419, -0.19731,
-1.12471, 1.50846, 1.18456, 0.50036, 0.61259, -0.95038, 3.26472, and 1.4098.

The sample mean and the standard deviation using the full data were 1.33 and 1.216, respectively. The following is the plot of the posterior density of the mean: $g^{*}(\mu \mid \sigma^{2}, x)$, using Mathematica. The posterior mean estimate using equation (18) is 1.359.

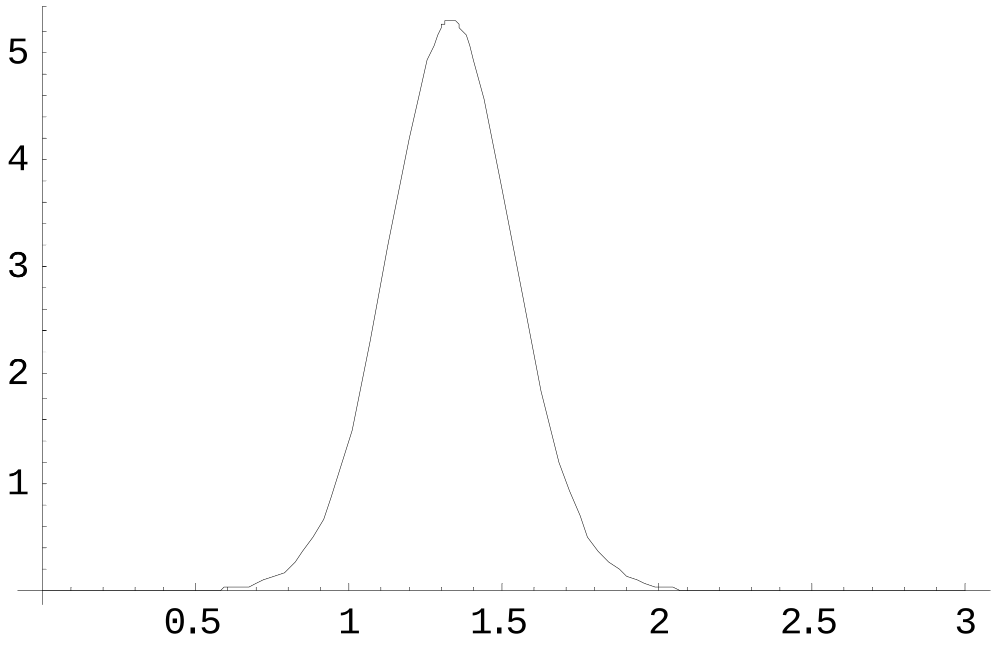


**Figure 11. Posterior Density of the mean,** $\mu = 1.33$, $\sigma = 1.216$

**Example 4B(Censored Data)**

A simulated data set of size 30 was generated from normal population with mean, $\mu$ =1 and $\sigma$ =1, N (1, 1) with L=.1 and k=4. The left censored data are: <.1, <.1, <.1, <.1, 0.11126, 1.75206, 1.17052, 3.15024, 3.18094, 1.56179, 0.927, 2.14169, 2.18118, 0.98145, 1.41042, 3.10198, 2.78779, 0.71599, 0.54362, 0.5441, 2.71058, 2.60982, 0.77772, 1.80419, 1.50846, 1.18456, 0.50036, 0.61259, 3.26472, 1.4098. The sample mean and the standard deviation obtained using the 26 observed values were 1.64 and 0.972, respectively. The following is the plot of the posterior density of the mean: $g^{*}(\mu \mid \sigma^{2}, x)$, using Mathematica. The posterior mean estimate using equation (14) is 1.7972 and the upper confidence level (UCL) is 1.82.

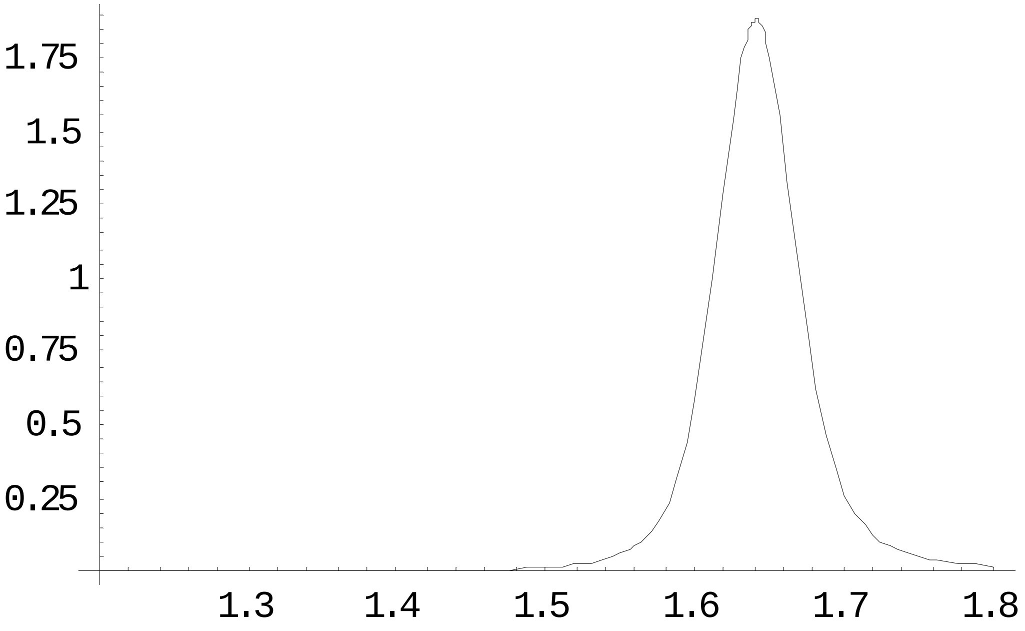


**Figure 12. Posterior Density of the mean,** $\mu = 1.64$, $\sigma = .972$

**Example 5A ( Full Data)**

A simulated data set of size 30 was generated from normal population with mean, $\mu$ =1 and $\sigma$ =1, N (1, 1), 0.98559, 1.72512, 0.76537, 3.06721, 3.1921, 2.01209, 1.91794, -0.11061, 0.96908, 1.09452, 2.52398, 0.4659, 1.60952, 0.54288, -0.17748, 0.74285, -0.32055,1.84595, -0.25303, 1.46591, 4.05311, 2.03353, -1.42666, -0.13452, -0.40622,0.88733, 1.35628, 1.36043, 2.10606, and 0.30362.

The sample mean and the standard deviation using the full data were 1.14 and 1.212, respectively. The following is the plot of the posterior density of the mean: $g^{*}(\mu \mid \sigma^{2}, x)$, using Mathematica. The posterior mean estimate using equation (18) is 1.196.

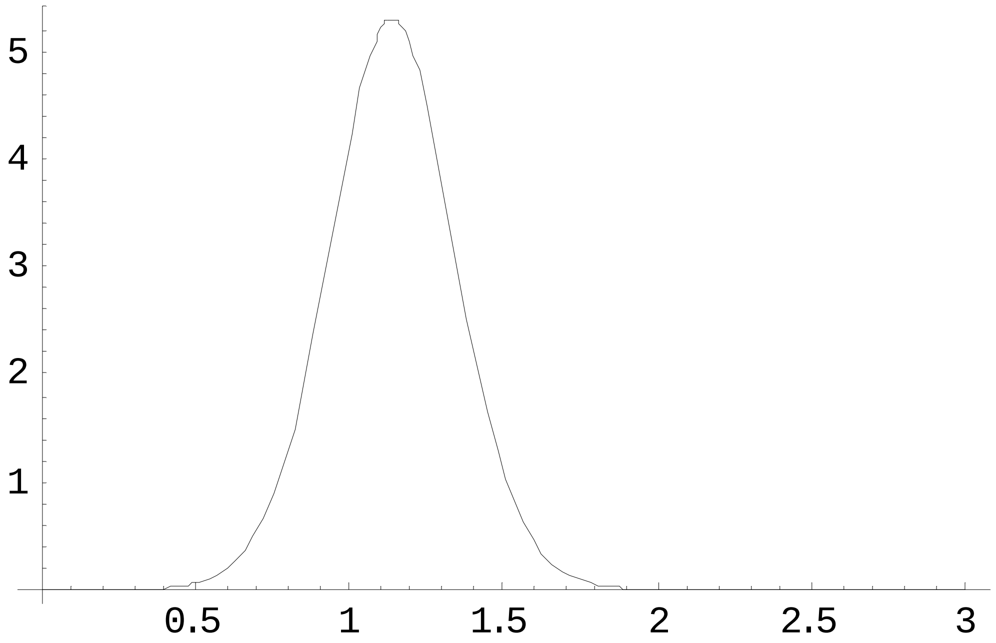


**Figure 13. Posterior Density of the mean,** $\mu = 1.41$, $\sigma = 1.212$

## Example 5B (Full Data)

A simulated data set of size 30 was generated from normal population with mean, $\mu$ =1 and $\sigma$ =1, N (1, 1) with L=.1 and k=7. The left censored data are: <.1, <.1, <.1, <.1, <.1, <.1, <.1, 0.98559, 1.72512, 0.76537, 3.06721, 3.1921, 2.01209, 1.91794, 0.96908, 1.09452, 2.52398, 0.4659, 1.60952, 0.54288, 0.74285, 1.84595, 1.46591, 4.05311, 2.03353, 0.88733, 1.35628, 1.36043, 2.10606, 0.30362. The sample mean and the standard deviation obtained using the 23 observed values were 1.610 and 0.942, respectively. The following is the plot of the posterior density of the mean: $g^*(\mu | \sigma^2, x)$, using Mathematica. The posterior mean estimate using equation (14) is 1.6825 and the upper confidence level (UCL) is 1.71.

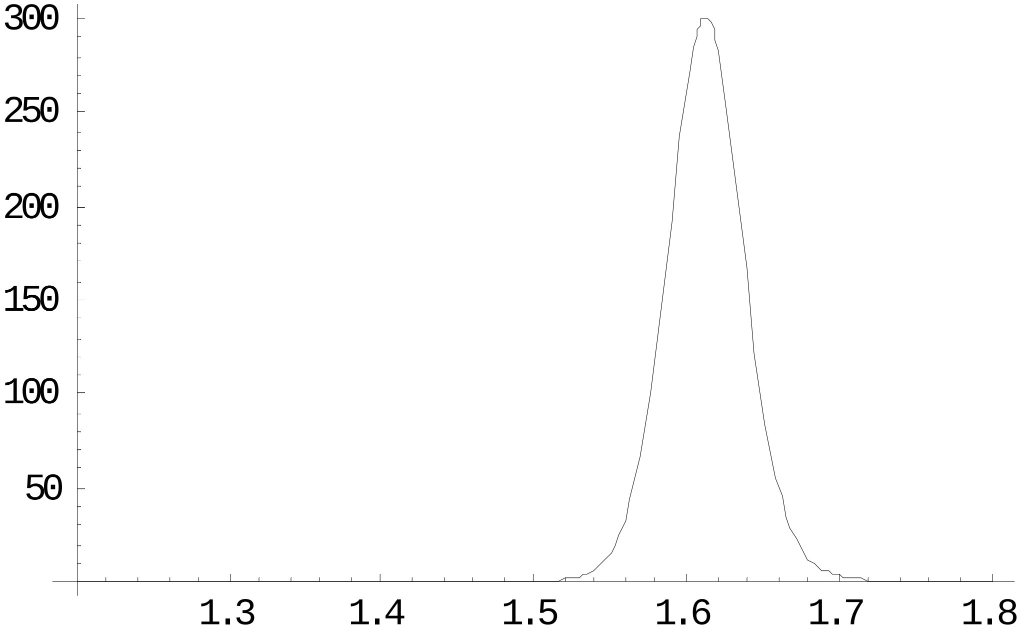


**Figure 14. Posterior Density of the mean,** $\mu = 1.610$, $\sigma = .942$

**Example 2 from Singh and Nocerino ( 2002)**

A simulated data set of size 15 was obtained from a normal population with mean, $\mu$ = 1.33 and standard deviation, $\sigma$ =.2, N(1.33,.2), with detection limit, L=1.0, and k=2. The left-censored data are: <1.0, <1.0, 1.2883, 1.1612, 1.156, 1.3251, 1.1568, 1.5638, 1.2914, 1.3253, 1.2884, 1.4688, 1.4581, 1.3641, and 1.1342. The sample mean and the standard deviation obtained from the 13 observed data values are 1.306 and 0.134, respectively. The following is the posterior probability density plot of the mean: $g^*(\mu | \sigma^2, x)$, of the left-censored data. The posterior mean estimate using equation (18) is 1.5123 and the upper confidence level (UCL) is 1.7.

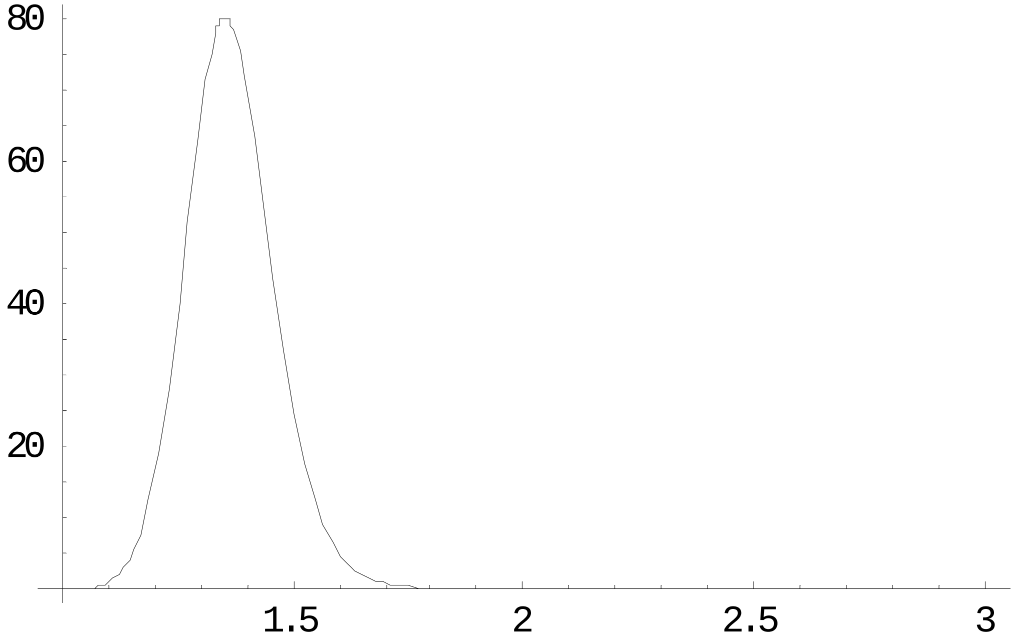


**Figure 15. Posterior Density of the mean,** $\mu = 1.306, \ \sigma = .134$

**Example 4 from Singh and Nocerino (2002)**

This left-censored data set is taken from the U.S. EPA RCRA guidance document [1992]. The detection limit, DL, is 1,450. The data has 3 non-detects and 21 observed values and they are: <1450, <1450, <1450, 1850, 1760, 1710, 1575, 1475, 1780, 1790, 1780, 1790, 1800, 1800, 1840, 1820, 1860, 1780, 1760, 1800, 1900, 1770, 1790, and 1780. The sample mean and standard deviation using the 21 observations are 1771.91 and 92.702, respectively. The following is the posterior probability density plot of the mean: $g^{*}(\mu \mid \sigma^{2}, x)$ of the left-censored data. The posterior mean estimate using equation (18) is 1771.7 and the upper confidence level (UCL) is 1775.

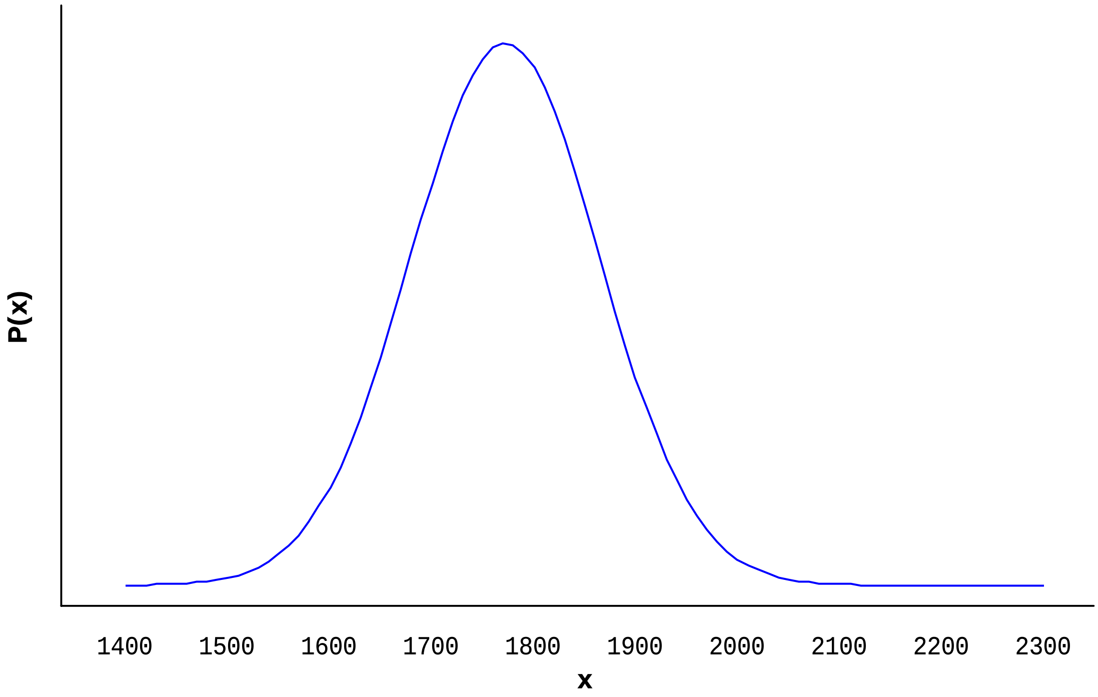


**Figure 16. Posterior Density of the mean,** $\mu = 1771.9,\ \sigma = 92.702$

**Table 1. Summary of the Data Simulation and Examples**

| Examples | Sample Size n | Sample Size Below DL | Detection Limit | Posterior Mean of Uncensored Data | Posterior Mean of Censored Data | 95%UCL of Censored Data | 95% UCL of uncensored data (bootstrap) | 95% UCL of censored data (bootstrap) |
|---|---|---|---|---|---|---|---|---|
| Ex1 | 30 | 4 | .1 | 1.307 | 1.5091 | 1.875 | 1.6 | 1.875 |
| Ex2 | 30 | 5 | .1 | .967 | 1.2565 | 1.33 | 1.138 | 1.429 |
| Ex3 | 30 | 2 | .1 | 1.307 | 1.5256 | 1.78 | 1.5 | 1.418 |
| Ex4 | 30 | 4 | .1 | 1.359 | 1.7972 | 1.82 | 1.697 | 1.933 |
| Ex5 | 30 | 7 | .1 | 1.196 | 1.6825 | 1.714 | 1.491 | 1.947 |
| Ex6 | 15 | 2 | 1 | N/A | 1.5123 | 1.70 | N/A | 1.369 |
| Ex7 | 24 | 3 | 1450 | N/A | 1771.7 | 1775 | N/A | 1797 |

# Section 5

## Summary and conclusion

In this paper, we have presented the most common method in dealing with left-censored data in Environmental application. Substitution and deletion methods assign arbitrary values between zero and DL or basically delete the censored data. These procedures result in loss of information and produce biased results. Another type of substitution is to replace the censored data by DL/2 or by the detection limit it self. However, this also results in biased estimate of the sample mean and variance if the censoring intensity is greater than 20%, Newman and Dixon (1990). Most of these substitution methods result in higher biased and larger MSE as sample size increases.

Several methods and examples to estimate the mean of left-censored data were mentioned such as the trimmed mean where the data set is sorted out by order and an equal number of observations can be trimmed, Gilbert (1987). The winsorized mean method was also recommended by Gilbert (1987) to estimate the mean of the left-censored data set and illustrated by an example in chapter 2. The maximum likelihood estimation was also one of the recommended methods but it will perform poorly for larger samples. The performances of these methods depend upon several things such as the sample size, the censoring intensity, and the value of the detection limits.

Bootstrap is one of the popular non-parametric procedures used in environmental applications and does not require assumption of the underlying distribution. Numerical examples show the estimate of the mean and the upper confidence limit (UCL) of the censored data and the uncensored data.

This paper mainly concerned with the Bayesian estimate of the mean of the left-censored data. Considering the non-informative prior, the posterior probability density function for left-censored data was obtained. However, this density function is not recognizable therefore; numerical integration was implemented to obtain the posterior mean and the upper confidence limit. Numerical examples results shown in table1 illustrate the various methods mentioned above.